FIRM FOUNDATION COUNTRY SCHOOL

Plot 8 Southwood, Chegutu, Zimbabwe
CONTACT: +263 77 643 4036/ +263 78 114 7858

CANDIDATE NAME

CANDIDATE CLASS

COMPUTER SCIENCE **0478**

Paper 1

HOLIDAY WORK TERM 1 2026

Form 2

You must answer on the question paper

INSTRUCTIONS

- Write your name in the space at the top of this page.
- Answer **all** questions.
- Use a black or dark blue pen.
- Write your answer to each question in the spaces provided.
- Do not use an erasable pen or correction fluid.
- Use a pencil for all drawings.
- Do not leave blank spaces.

DO NOT WRITE IN ANY OTHER SPACES.

INFORMATION

- You are required to print this document.
- The number of marks for each question or part question is shown in brackets [ ].

This document has 126 pages

1 Parity checks are often used to check for errors that may occur during data transmission.

(a) A system uses **even parity**.

Tick (✓) to show whether the following three bytes have been transmitted correctly or incorrectly.

Received byte	Byte transmitted correctly	Byte transmitted incorrectly
1 1 0 0 1 0 0 0		
0 1 1 1 1 1 0 0		
0 1 1 0 1 0 0 1		

[3]

(b) A parity byte is used to identify which bit has been transmitted incorrectly in a block of data.

The word "F L O W C H A R T" was transmitted using nine bytes of data (one byte per character). A tenth byte, the parity byte, was also transmitted.

The following block of data shows all ten bytes received after transmission. The system uses **even parity** and column 1 is the parity bit.

	letter	column 1	column 2	column 3	column 4	column 5	column 6	column 7	column 8
byte 1	F	1	0	1	0	0	1	1	0
byte 2	L	1	0	1	0	1	1	0	0
byte 3	O	1	0	1	0	1	1	1	1
byte 4	W	1	0	1	1	0	1	1	1
byte 5	C	1	0	1	0	0	0	1	1
byte 6	H	0	0	1	0	1	0	0	0
byte 7	A	0	0	1	0	0	1	0	1
byte 8	R	1	0	1	1	0	0	1	0
byte 9	T	1	0	1	1	0	1	0	0
parity byte		1	0	1	1	1	1	1	0

(i) **One** of the bits has been transmitted incorrectly.

Write the byte number and column number of this bit:

Byte number .....

Column number .....

[2]

(ii) Explain how you arrived at your answer for **part (b)(i)**.

.....

.....

.....

.....[2]

(c) Give the denary (base 10) value of the byte: **1 0 1 1 1 1 1 0**

.....

.....[1]

(d) A parity check may not identify that a bit has been transmitted incorrectly.

Describe **one** situation in which this could occur.

.....

.....

.....[1]

2 **(a)** Nikita wishes to print out some documents and connects her printer to the computer using one of the USB ports.

(i) Identify what type of data transmission is being used.

.....[1]

(ii) Give **three** reasons for using a USB port.

1 .....

.....

2 .....

.....

3 .....

.....

[3]

(iii) The printer runs out of paper while it is printing the documents. A signal is sent to the processor to request that the problem is dealt with.

Name this type of signal.

.....[1]

(b) State **one** suitable application for **each** printer below. A different application must be given for each printer.

Inkjet printer .....

.....

3D printer .....

.....

[2]

(c) Name another type of printer and describe **one** way in which it is different from the printers named in **part (b)**.

Give an application for this printer.

Type of printer .....

Description .....

.....

.....

Application .....

.....

[3]

3 **(a)** Parity checks are often used to detect errors that may occur during data transmission.

The received bytes in the table below were transmitted using **odd parity**.

Tick (✓) to show whether each byte has been **corrupted during transmission** or **not corrupted during transmission**.

Received byte	corrupted during transmission (✓)	not corrupted during transmission (✓)
10110100		
01101101		
10000001		

[3]

(b) Another method of error detection is Automatic Repeat reQuest (ARQ).

Explain how ARQ is used in error detection.

.....

.....

.....

.....

.....

.....

.....

.....

.....

..... [4]

4 Computer A is communicating with computer B.

(a) Draw an arrow or arrows to show simplex, duplex and half-duplex data transmission. The **direction** of the data transmission must be fully **labelled**.

Simplex data transmission

Computer **A**

Computer **B**

Duplex data transmission

Computer **A**

Computer **B**

Half-duplex data transmission

Computer **A**

Computer **B**

[6]

(b) State a use for the following data transmission methods. The use must be different for each data transmission method.

Simplex .....

.....

Duplex .....

.....

[2]

(c) A computer includes an Integrated Circuit (IC) and a Universal Serial Bus (USB) for data transmission.

Describe how the computer uses these for data transmission, including the type of data transmission used.

IC .....

.....

.....

.....

USB .....

.....

.....

.....

[4]

5 RockICT is a music business that has a website to allow customers to view and buy the products it sells.

The website consists of web pages.

(a) Describe what is meant by HTML structure and presentation for a web page.

.....

.....

.....

.....

.....

.....

.....

.....[4]

(b) The URL for the music company's website is:

https://www.rockict.net/index.htm

Part 1 **Part 2**

(i) Identify what **Part 1** and **Part 2** represent in this URL.

Part 1 .....

Part 2 .....
[2]

(ii) Describe what is meant by **https**.

.....

.....

.....

.....[2]

(c) When a customer enters the website, a message is displayed:

"RockICT makes use of cookies. By continuing to browse you are agreeing to our use of cookies."

Explain why the music company uses cookies.

.....

.....

.....

.....

[2]

(d) The music company is concerned about the security of its website.

The company uses a proxy server as part of its security system.

Describe the role of a proxy server in the security system.

.....

.....

.....

.....

.....

.....

.....

.....

[4]

6 A company transmits data to external storage at the end of each day.

(a) Parity checks can be used to check for errors during data transmission.

The system uses **odd parity**.

(i) Tick (✓) to show for each of the received bytes whether they have been **transmitted correctly** or **transmitted incorrectly**.

Received byte	**Transmitted correctly (✓)**	**Transmitted incorrectly (✓)**
10001011		
10101110		
01011101		
00100101		

[4]

(ii) State **one** other method that could be used to check for transmission errors.

..... [1]

(b) Data can be transferred using parallel or serial data transmission.

(i) Describe what is meant by parallel data transmission.

.....

.....

.....

..... [2]

(ii) Give **one** application of parallel data transmission.

.....

..... [1]

(iii) Explain why serial data transmission is normally used for transferring data over a long distance.

.....

.....

.....

..... [2]

(c) Data transferred over a network is encrypted to improve data security.

The system uses 64-bit symmetric encryption.

(i) Explain how encryption improves data security.

.....

.....

.....

..... [2]

(ii) Explain **one** method that could be used to increase the level of security provided by the encryption.

.....

.....

.....

..... [2]

7 Leonard has a new laser printer to print letters for his business.

Leonard connects his printer to his computer using the USB port.

(a) Give **three** benefits of using the USB port to connect the printer to the computer.

Benefit 1 .....

.....

Benefit 2 .....

.....

Benefit 3 .....

.....

[3]

(b) State **two** benefits and **one** drawback of Leonard using a laser printer, instead of an inkjet printer, to print the letters.

Benefit 1 .....

.....

Benefit 2 .....

.....

Drawback .....

.....

[3]

(c) An interrupt signal is sent from the printer to the computer.

(i) Give **two** examples of when a printer would generate an interrupt signal.

Example 1 .....

Example 2 .....

[2]

(ii) Many devices send interrupt signals.

Identify the software in the computer that will receive and manage all interrupt signals.

..... [1]

8 Four 7-bit binary values are being transmitted from one computer to another. An odd parity check is being used to check for errors in the binary values.

Write the correct **Parity bit** for each **7-bit binary value** to make sure it meets **odd** parity.

Parity bit	**7-bit binary value**
.....	0000011
.....	1000000
.....	0111111
.....	1010101

[4]

9 Clive has a laptop computer that he uses for his business. He enters a username and password to log in to his laptop.

Clive is worried about spyware being used to find out his username and password.

(a) Describe how spyware could be used to find out Clive's username and password.

.....

.....

.....

.....

.....

.....

.....

..... [4]

(b) The threat of spyware makes Clive concerned about typing a password to log in to his laptop.

Give an example of how Clive could log in securely without typing a password.

..... [1]

10 Four descriptions about compilers and interpreters are shown below.

Draw lines to indicate which descriptions refer to a compiler and which descriptions refer to an interpreter.

Description

It is more difficult to debug the code since one error can produce many other associated errors.

The speed of execution of program loops is slower.

It produces fast, executable code that runs directly on the processor.

It is easier to debug the code since an error is displayed as soon as it is found.

Compiler

Interpreter

[4]

11 A file server is used as a central data store for a network of computers.

Rory sends data from his computer to a file server that is approximately 100 metres away.

It is important that the data is transmitted accurately. Rory needs to be able to read data from and write data to the file server at the same time.

(a) (i) Use ticks (✓) to identify the most suitable data transmission methods for this application.

Method 1	Tick (✓)	**Method 2**	Tick (✓)
Serial		Simplex	
Parallel		Half-duplex	
		Duplex	

[2]

(ii) Explain why your answer to **part (a)(i)** is the most suitable data transmission.

.....

.....

.....

.....

.....

.....

.....

.....[4]

(b) Identify and describe **two** methods of error checking that can be used to make sure that the data stored after transmission is accurate.

Method 1 .....

.....

.....

.....

.....

.....

.....

.....

Method 2 .....

.....

.....

.....

.....

.....

.....

.....

[6]

12 Kamil correctly answers an examination question about a number of internet terms.

Six different terms have been removed from Kamil's answer.

Complete the sentences in Kamil's answer, using the list given. Not all terms in the list need to be used.

- browser
- connection
- domain name server (DNS)
- Internet
- Internet Service Provider (ISP)
- IP address
- MAC address
- network
- protocol
- uniform resource locator (URL)
- webpages
- hypertext mark-up language (HTML)

A ..... is a program that allows a user

to view ..... .

An ..... is a company that provides a

connection to access the ..... .

The main ..... that governs the

transmission of data using the Internet is http.

The ..... is provided by the network,

and given to each device on the network.

[6]

13 **(a)** Computers can transmit data using different methods.

Describe the **three** data transmission methods given.

(i) Serial data transmission

.....

.....

.....

.....[2]

(ii) Parallel data transmission

.....

.....

.....

.....[2]

(iii) Duplex data transmission

.....

.....

.....

.....[2]

(b) Data can sometimes be corrupted when it is transmitted from one computer to another, causing errors to be present in the data.

Identify and describe **three** methods of error detection that could be used to see if an error has occurred.

Error detection method 1 .....

Description .....

.....

.....

.....

Error detection method 2 .....

Description .....

.....

.....

.....

Error detection method 3 .....

Description.....

.....

.....

.....

[9]

14 The MAC address of a device is represented using hexadecimal.

A section of a MAC address is shown. Each pair of hexadecimal digits is stored using 8-bit binary.

(a) Complete the table to show the 8-bit binary equivalents for the section of MAC address. The first number has already been converted.

6A	FF	08	93
01101010			

[3]

(b) Explain why data is stored as binary in computers.

.....

.....

.....

.....[2]

15 Data can be transferred using half-duplex serial transmission.

(a) Describe serial transmission.

.....

.....

.....

.....[2]

(b) Give **one** application of serial data transmission.

.....

.....[1]

(c) Describe half-duplex data transmission.

.....

.....

.....

.....[2]

16 Sonia shares files with her friends. The method of data transmission she uses is half-duplex serial transmission.

(a) Describe how data is transmitted using half-duplex serial data transmission.

.....

.....

.....

.....

.....

.....

.....

..... [4]

(b) The system uses parity bits to check for errors during data transmission.

The outcome of four bytes after transmission is:

Byte 1	Byte 2	Byte 3	Byte 4
00110011	01010100	10110100	01110111

One of the bytes has been transmitted incorrectly.

Identify the byte that was transmitted incorrectly.

Byte .....

Explain how you identified the byte that was transmitted incorrectly.

.....

.....

.....

.....

[4]

17 Data is valuable to a company.

(a) Companies use error detection methods to make sure that data is accurate.

One error detection method is the use of a check digit.

Explain what is meant by a check digit and how it is used to detect errors.

.....

.....

.....

.....

.....

.....

.....

..... [4]

(b) Companies can use a range of security methods to keep their data secure.

Identify **two** security methods that a company can use to keep their data secure **and** explain how each method can keep the data secure.

Security method 1 .....

.....

.....

.....

.....

.....

Security method 2 .....

.....

.....

.....

.....

.....

[6]

18 The contents of three binary registers have been transmitted from one computer to another. **Odd parity** has been used as an error detection method.

The outcome after transmission is:

- **Register A** and **Register B** have been transmitted **correctly**.
- **Register C** has been transmitted **incorrectly**.

Write the appropriate **Parity bit** for each register to show the given outcome.

	Parity bit							
Register A		0	1	0	0	0	1	1
Register B		0	0	0	0	1	1	1
Register C		0	0	0	0	0	1	1

[3]

19 Jesse is taking his Computer Science examination. He answers **five** questions about ethics.

(a) For the first question, he writes the answer:

"This type of software can be copied and shared without the permission of the owner."

State what Jesse is describing.

..... [1]

(b) For the second question, he writes the answer:

"With this type of software, the owner still retains the copyright for the software, but he gives away copies of it for free."

State what Jesse is describing.

..... [1]

(c) For the third question, he writes the answer:

"This type of software is often a trial version of the full software. To use the full version the user normally needs to pay a fee."

State what Jesse is describing.

..... [1]

2 Data transmission

In this chapter you will learn about:

- ★ types and methods of data transmission
  - how data is broken up into data packets before transmission
  - the structure of data packets (header, payload and trailer)
  - packet switching (including the role of the router in the process)
  - methods of data transmission (serial, parallel, simplex, half-duplex and full-duplex)
  - the Universal Serial Bus (USB)
- ★ methods of error detection
  - why error checking methods are needed
  - error checking methods following data transmission:
    - parity checks
    - checksum
    - echo check
  - use of check digits to detect data entry errors
  - use of automatic repeat requests (ARQs) to detect errors
- ★ encryption
  - the need for and the purpose of encryption
  - symmetric and asymmetric encryption
  - use of public and private keys.

Data is frequently transferred from one device to another. The two devices could be in the same building or thousands of kilometres away. Irrespective of the distance travelled, the transmission of data needs to be considered with respect to:

- how the data is transmitted
- how can errors following transmission be detected and can the data be recovered
- the role of encryption to make sure data that falls into the wrong hands can't be used

It is also important to consider ways of checking for errors in data once it has been entered into a computer.

2.1 Types and methods of data transmission

2.1.1 Data packets

Data packets are usually referred to simply as 'packets'

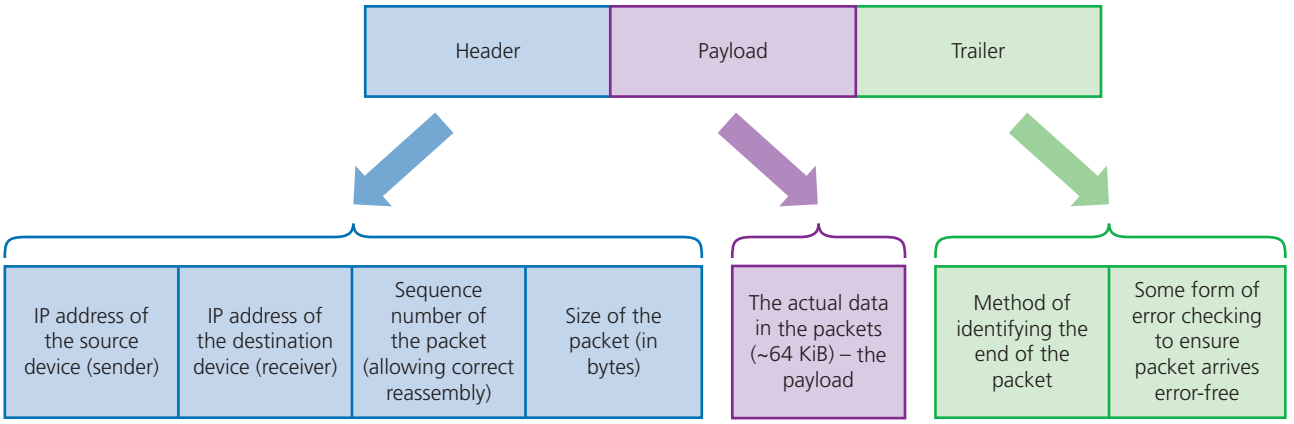
Data sent over long distances is usually broken up into **data packets** (sometimes called datagrams). The packets of data are usually quite small, typically 64 KiB, which are much easier to control than a long continuous stream of data. The idea of splitting up data in this way means each packet can be sent along a different route to its destination. This would clearly be of great benefit if a particular transmission route was out of action or very busy. The only obvious drawback of

splitting data into packets is the need to reassemble the data when it reaches its destination.

Packet structure

A typical packet is split up into:

- a packet header
- the payload
- a trailer.



▲ **Figure 2.1** Packet structure

For **each** packet, the **packet header** consists of:

- the IP address of the sending device
- the IP address of the receiving device
- the sequence number of the packet (this is to ensure that all the packets can be reassembled into the correct order once they reach the destination)
- packet size (this is to ensure the receiving station can check if all of the packets have arrived intact).

(Note: the header often also contains another value indicating how many packets there are in total for this transmission.)

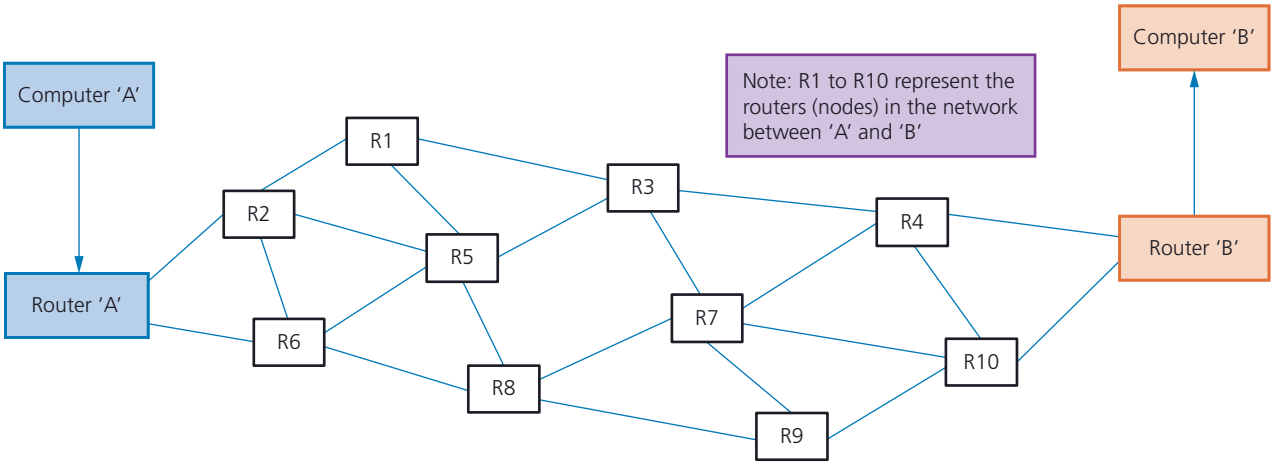
For **each** packet, the **payload** consists of the actual data being sent in the packet (this is usually about 64 KiB).

For **each** packet, the **packet trailer** consists of:

- some way of identifying the end of the packet; this is essential to allow each packet to be separated from each other as they travel from sending to receiving station
- an error checking method; **cyclic redundancy checks** (CRCs) are used to check data packets:
  - this involves the sending computer adding up all the 1-bits in the payload and storing this as a hex value in the trailer before it is sent
  - once the packet arrives, the receiving computer recalculates the number of 1-bits in the payload
  - the computer then checks this value against the one sent in the trailer
  - if the two values match, then no transmission errors have occurred; otherwise the packet needs to be re-sent.

Packet switching

Let us now consider what happens when a photograph, for example, is sent from computer 'A' to computer 'B'. The photograph will be split up into a number of packets before it is sent. There will be several possible routes for the packets, between computer 'A' (sender) and computer 'B' (receiver). Each stage in the route contains a **router**. A router receives a data packet and, based on the information in the header, decides where to send it next. For example:



▲ **Figure 2.2** Typical network showing possible routes between 'A' and 'B'

Packet switching is a method of data transmission in which a message is broken up into a number of packets. **Each** packet can then be sent independently from start point to end point. At the destination, the packets will need to be reassembled into their correct order (using the information sent in the header). At each stage in the transmission, there are **nodes** that contain a router. Each router will determine which route the packet needs to take, in order to reach its destination (the destination IP address is used in this part of the process).

Suppose our photograph (Figure 2.3) has been split up into five packets that have been sent in the following order:

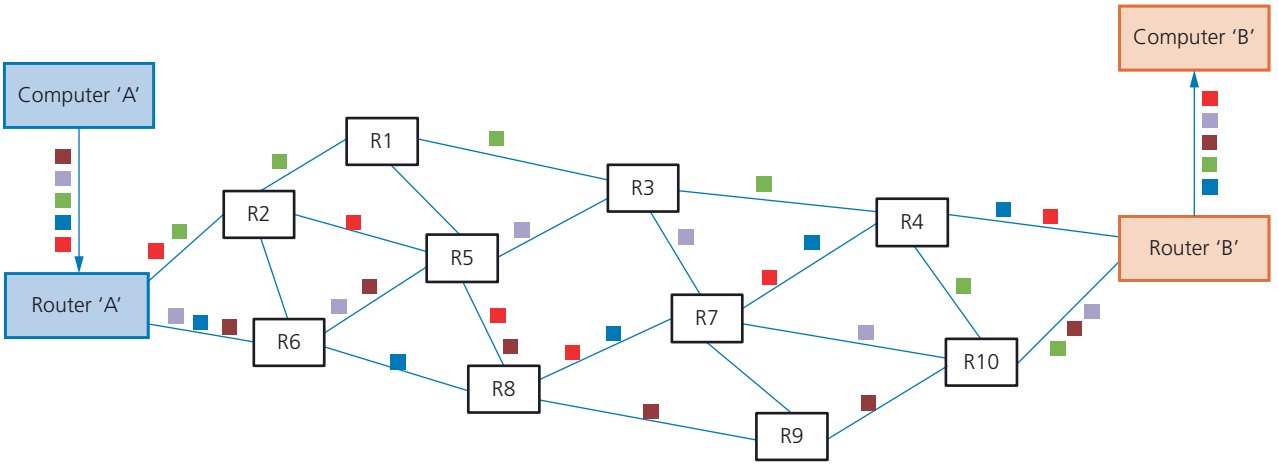
▲ **Figure 2.3**

- each packet will follow its own path (route)
- routers will determine the route of each packet
- routing selection depends on the number of packets waiting to be processed at each node
- the shortest possible path **available** is always selected – this may not always be the shortest path that **could** be taken, since certain parts of the route may be too busy or not suitable
- unfortunately, packets can reach the destination in a different order to that in which they were sent.

Figure 2.5 shows one possible scenario. Notice the different paths taken by each packet from computer 'A' to computer 'B'. Also notice that the packets have arrived in a different order compared to the way they were sent, namely:

▲ **Figure 2.4**

Computer 'B' will now have to reassemble the packets into the original sequence.



▲ **Figure 2.5** Typical network showing possible paths taken by each packet

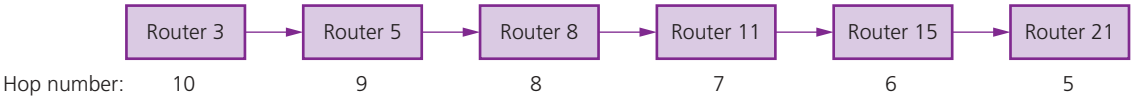
The benefits of packet switching are:

- there is no need to tie up a single communication line
- it is possible to overcome failed, busy or faulty lines by simply re-routing packets
- it is relatively easy to expand package usage
- a high data transmission rate is possible.

The drawbacks of packet switching include:

- packets can be lost and need to be re-sent
- the method is more prone to errors with **real-time streaming** (for example, a live sporting event being transmitted over the internet)
- there is a delay at the destination whilst the packets are being re-ordered.

Sometimes it is possible for packets to get lost because they keep 'bouncing' around from router to router and never actually reach their destination. Eventually the network would just grind to a halt as the number of **lost packets** mount up, clogging up the system. To overcome this, a method called **hopping** is used. A **hop number** is added to the header of each packet, and this number is reduced by 1 *every time it leaves a router* (Figure 2.6).



▲ **Figure 2.6** Hop numbers between routers

Each packet has a maximum hop number to start with. Once a hop number reaches zero, and the packet hasn't reached its destination, then the packet is deleted when it reaches the next router. The missing packets will then be flagged by the receiving computer and a request to re-send these packets will be made.

Find out more

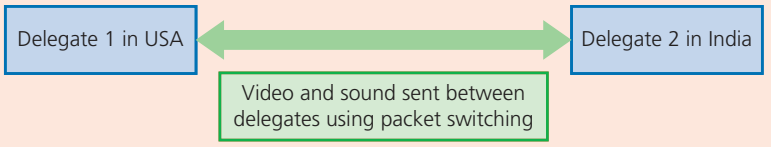
Another method of sending packets is called circuit switching. Find out how this differs to packet switching, and then re-draw Figure 2.5 showing the route the packets take when using circuit switching.

Advice

Hopping is not included on the syllabus but is included here for completeness (to help understand how packets can get lost).

Activity 2.1

1 Suppose a video conference is taking place between delegates in two different countries. Packet switching is being used to send video and sound data between the delegates:



▲ **Figure 2.7**

Describe:

i any potential problems with sound and video quality

ii how these problems could be caused.

2 Explain how packet switching could be used to download a large web page from a website.

3 a The trailer in a packet will use one form of error checking.

Explain what is meant by a cyclic redundancy check.

b The payload contains the following data:

```
11110000 10000011 00110011 00111111 11111110 11100011
```

Use this data to show how the receiving computer can verify that the received payload was error-free.

4 a Explain how it is possible for packets to be lost during their transmission across a network.

b Describe how it is possible for a system to deal with lost packets and prevent them from slowing down the transmission process.

c Explain why you think packet switching might improve data security.

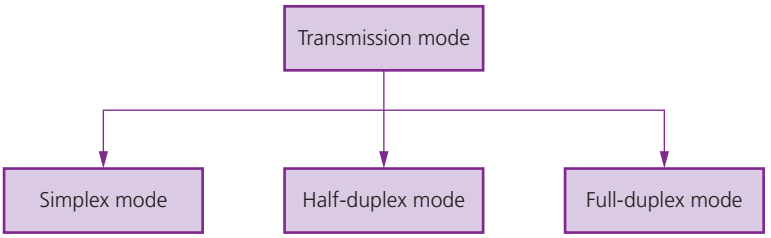
2.1.2 Data transmission

Data transmission can be either over a short distance (for example, computer to printer) or over longer distances (for example, from one computer to another in a global network). Essentially, three factors need to be considered when transmitting data:

- the direction of data transmission (for example, can data transmit in one direction only, or in both directions)
- the method of transmission (for example, how many bits can be sent at the same time)
- how will data be synchronised (that is, how to make sure the received data is in the correct order).

These factors are usually considered by a communication protocol.

Simplex, half-duplex and full-duplex



▲ **Figure 2.8** Transmission modes

Simplex data transmission

Simplex mode occurs when data can be sent in **ONE DIRECTION ONLY** (for example, from sender to receiver). An example of this would be sending data from a computer to a printer.

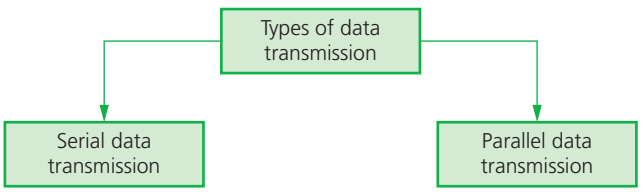
Half-duplex data transmission

Half-duplex mode occurs when data is sent in **BOTH DIRECTIONS** but **NOT AT THE SAME TIME** (for example, data can be sent from 'A' to 'B' and from 'B' to 'A' along the same transmission line, but they can't both be done at the same time). An example of this would be a walkie-talkie where a message can be sent in one direction only at a time; but messages can be both received and sent.

Full-duplex data transmission

Full-duplex mode occurs when data can be sent in **BOTH DIRECTIONS AT THE SAME TIME** (for example, data can be sent from 'A' to 'B' and from 'B' to 'A' along the same transmission line simultaneously). An example of this would be a broadband internet connection.

Serial and parallel data transmission



▲ **Figure 2.9** Types of data transmission

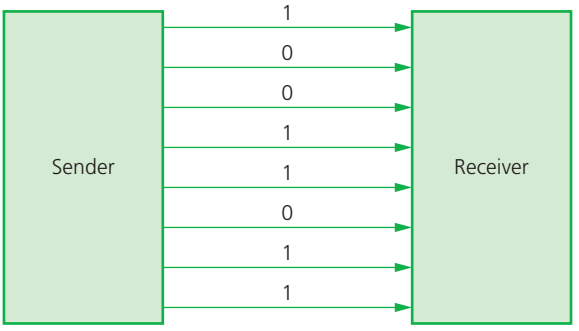
Serial data transmission occurs when data is sent **ONE BIT AT A TIME** over a **SINGLE WIRE/CHANNEL**. Bits are sent one after the other as a single stream.

▲ **Figure 2.10** Serial data transmission

(Note: Serial data transmission can be simplex, half-duplex or full-duplex.)

Serial data transmission works well over long distances. However, the data is transmitted at a slower rate than parallel data transmission. Because only one channel/wire is used, data will arrive at its destination fully synchronised (i.e. in the correct order). An example of its use is when connecting a computer to a printer via a USB connection (see Section 2.1.3).

Parallel data transmission occurs when **SEVERAL BITS OF DATA** (usually one byte) are sent down **SEVERAL CHANNELS/WIRES** all at the same time. Each channel/wire transmits one bit:



▲ **Figure 2.11** Parallel data transmission

(Note: Parallel data transmission can be simplex, half-duplex or full-duplex.)

Parallel data transmission works well over short distances. Over longer distances (for example, over 20 metres), data can become **skewed** (that is, the data can arrive unsynchronised) and bits can arrive out of order. The longer the wire, the worse this can become. It is, however, a faster method of data transmission than serial. The internal circuits in a computer use parallel data transmission since the distance travelled between components is very short and high-speed transmission is essential.

Link

For more on data transmission within the CPU refer to Chapter 3.

Activity 2.2

1 Explain what is meant by:
   - i serial, half-duplex data transmission
   - ii parallel, full-duplex data transmission
   - iii serial, simplex data transmission.

2 Which types of data transmission are being described:
   - i data is sent one bit at a time in one direction only
   - ii data is being sent 8 bits at a time in one direction only
   - iii data is being sent 16 bits at a time in both directions simultaneously
   - iv data is sent one bit at a time in both directions simultaneously
   - v data is sent 16 bits at a time in one direction only?

Table 2.1 shows the comparison between serial and parallel data transmission.

▼ **Table 2.1** Comparison of serial and parallel data transmission methods

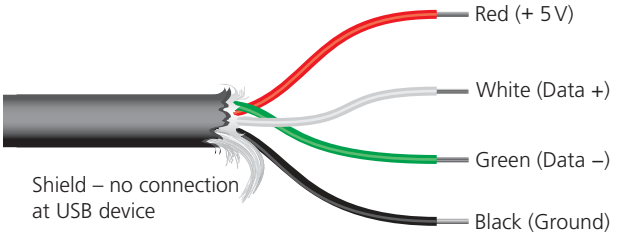
Serial	Parallel
less risk of external interference than with parallel (due to fewer wires)	faster rate of data transmission than serial
more reliable transmission over longer distances	works well over shorter distances (for example, used in internal pathways on computer circuit boards)
transmitted bits won't have the risk of being skewed (that is, out of synchronisation)	since several channels/wires used to transmit data, the bits can arrive out of synchronisation (skewed)
used if the amount of data being sent is relatively small since transmission rate is slower than parallel (for example, USB uses this method of data transmission)	preferred method when speed is important
used to send data over long distances (for example, telephone lines)	if data is time-sensitive, parallel is the most appropriate transmission method
less expensive than parallel due to fewer hardware requirements	parallel ports require more hardware, making them more expensive to implement than serial ports
	easier to program input/output operations when parallel used

2.1.3 Universal serial bus (USB)

Link

Also refer to Section 4.1 on software drivers regarding devices plugged into USB ports.

As the name suggests, the **universal serial bus (USB)** is a form of serial data transmission. USB is now the most common type of input/output port found on computers and has led to a standardisation method for the transfer of data between devices and a computer. It is important to note that USB allows both half-duplex and full-duplex data transmission.



▲ **Figure 2.12** Typical USB cable

As Figure 2.12 shows, the USB cable consists of a four-wired shielded cable, with two wires for power (red and black). The other two wires (white and green) are for data transmission. When a device is plugged into a computer using one of the USB ports:

- the computer automatically detects that a device is present (this is due to a small change in the voltage on the data signal wires in the USB cable)
- the device is automatically recognised, and the appropriate device driver software is loaded up so that the computer and device can communicate effectively
- if a new device is detected, the computer will look for the device driver that matches the device; if this is not available, the user is prompted to download the appropriate driver software (some systems do this automatically and the user will see a notice asking for permission to connect to the device website).

We will now consider the benefits and drawbacks of using the USB system:

▼ **Table 2.2** Benefits and drawbacks of USB systems

Benefits	Drawbacks
devices plugged into the computer are automatically detected and device drivers are automatically loaded up	standard USB only supports a maximum cable length of 5 m; beyond that, USB hubs are needed to extend the cable length
connections can only fit one way preventing incorrect connections being made	
it has become an industry standard, which means considerable support is available	even though USB is backward compatible, very early USB standards (V1) may not always be supported by the latest computers
can support different data transmission rates (from 1.5 Mbps to 5 Gbps)	
no need for external power source since cable supplies +5 V power	even the latest version 3 (V3) and version 4 (V4) USB-C systems have a data transfer rate which is slow compared to, for example, Ethernet connections (Note: USB V2 has a maximum data transfer rate of 480 Mbps.)
USB protocol notifies the transmitter to re- transmit data if any errors are detected; this leads to error-free data transmission	
it is relatively easy to add more USB ports if necessary, by using USB hubs	
USB is backward compatible (that is, older versions are still supported)	

A new type of USB connector, referred to as USB-C, is now becoming more common in laptops and tablets/phones. This is a 24-pin symmetrical connector which means it will fit into a USB-C port either way round. It is much smaller and thinner than older USB connectors, offers 100 watt (20 volt) power connectivity, which means full-sized devices can now be charged and it can carry data at 10 gigabits per second (10 Gbps); this means it can now support 4K video delivery.

USB-C is backward compatible (to USB 2.0 and 3.0) provided a suitable adaptor is used, and is expected to become the new industry standard (universal) format.

Activity 2.3

Ten statements are shown on the left in the table. Each of these statements is either True or False. For each statement, tick (✓) the appropriate column to indicate which statements are true and which are false.

Statement	True	False
Packets have a header which contains the IP address of the sender and the receiver		
Packets don't require any form of error checking		
USBs use a protocol that allows for error-free data transmission between device and computer		
Serial data transmission suffers from data skewing		
The longest cable length supported by USB is 5 metres or less		
Simplex data transmission occurs when data is transmitted one bit at a time		
Full-duplex data transmission involves sending 8 bits of data at a time		
USB uses serial data transfer		
Packet switching prevents loss of any data packets		
USB connections can transfer data using half-duplex or full-duplex		

2.2 Methods of error detection

2.2.1 The need to check for errors

When data is transmitted, there is always a risk that it may be corrupted, lost or even gained.

Errors can occur during data transmission due to:

- interference (all types of cable can suffer from electrical interference, which can cause data to be corrupted or even lost)
- problems during packet switching (this can lead to data loss – or it is even possible to gain data!)
- skewing of data (this occurs during parallel data transmission and can cause data corruption if the bits arrive out of synchronisation).

Checking for errors is important since computers are unable to understand text, for example, if the words are not recognised by its built-in dictionary. Look at the following example of some corrupted text:

Can you raed tihs?

"I cnduo't bvleiee taht I culod aulaclty uesdtannrd waht I was rdnaieg. Unisg the icndeblire pweor of the hmuan mnid, aocdcrnig to rseecrah at Cmabridge Uinervtisy, it dseno't mttaer in waht oderr the lterets in a wrod are, the olny irpoamtnt tihng is taht the frsit and lsat ltteer be in the rhgit pclae. The rset can be a taotl mses and you can sitll raed it whoutit a pboerlm.

Tihs is bucseae the huamn mnid deos not raed ervey ltteer by istlef, but the wrod as a wlohe.

Aaznmig, huh? Yeah and I awlyas tghhuot slelinpg was ipmorantt! See if yuor fdreins can raed tihs too"

(from an unknown source at Cambridge University)

▲ **Figure 2.13** Example of data corruption on a message

Whilst you probably had little problem understanding this text, a computer would be unable to make any sense of it. Data corruption is therefore a very real problem to a computer. Figure 2.13 could be the result of some data corruption following transmission which would make the text unintelligible to a computer. This is why error checking is such an important part of computer technology. The following section considers a number of ways that can be used to check for errors, so that you don't end up with text as shown in Figure 2.13 above!

There are a number of ways data can be checked for errors following transmission:

- parity checks
- checksum
- echo check.

2.2.2 Parity checks, checksum and echo checks

Parity checks

Parity checking is one method used to check whether data has been changed or corrupted following data transmission. This method is based on the number of 1-bits in a byte of data.

The parity can be either called **EVEN** (that is, an even number of 1-bits in the byte) or **ODD** (that is, an odd number of 1-bits in the byte). One of the bits in the byte (usually the most significant bit or left-most bit) is reserved for a **parity bit**. The parity bit is set according to whether the parity being used is even or odd. For example, consider the byte:

	1	1	0	1	1	0	0

parity bit

Advice

Note for the 7 bit number 1110000, the even parity bit would be 1 and the odd parity bit would be 0. The parity bit can be set as a 1 or a 0 for either choice of parity – it just depends on how many 1s are in the byte.

In this example, if the byte is using even parity, then the parity bit needs to be set to 0, since there is already an even number of 1-bits in the byte (four 1-bits). We thus get:

0	1	1	0	1	1	0	0

parity bit

In this example, if the byte is using odd parity, then the parity bit needs to be set to 1, since we need to have an odd number of 1-bits in the byte. We thus get:

1	1	1	0	1	1	0	0

parity bit

Before data is transferred, an agreement is made between sender and receiver regarding which type of parity is being used. Parity checks are therefore being used as a type of transmission protocol.

Activity 2.4

Find the parity bits for each of the following bytes:

	1	1	0	1	1	0	1	even parity being used
	0	0	0	1	1	1	1	even parity being used
	0	1	1	1	0	0	0	even parity being used
	1	1	1	0	1	0	0	odd parity being used
	1	0	1	1	0	1	1	odd parity being used
	1	1	1	1	1	1	0	even parity being used
	1	1	1	1	1	1	0	odd parity being used
	1	1	0	1	0	0	0	odd parity being used
	0	0	0	0	1	1	1	even parity being used
	1	1	1	1	1	1	1	odd parity being used

If a byte has been transmitted from 'A' to 'B', and if even parity is used, an error would be flagged if the byte now had an odd number of 1-bits at the receiver's end. For example (assuming even parity is being used):

byte being sent:	0	1	0	1	1	1	0	0
	parity bit							
byte being received:	0	1	0	0	1	1	0	0
	parity bit							

In this case, the byte received has three 1-bits, which means it now has odd parity; while the sender's byte was using even parity (four 1-bits). This means an error has occurred during the transmission of the byte. The error is detected by the recipient's computer re-calculating the parity of the byte sent. If even parity had been agreed between sender and receiver, then a change in parity in the received byte indicates that a transmission error has occurred.

Activity 2.5

1 Which of the following received bytes indicate an error has occurred following data transmission?

1	1	1	0	1	1	0	1	even parity being used
0	1	0	0	1	1	1	1	even parity being used
0	0	1	1	1	0	0	0	even parity being used
1	1	1	1	0	1	0	0	odd parity being used
1	1	0	1	1	0	1	1	odd parity being used
1	1	1	1	1	1	1	1	odd parity being used
0	0	0	0	0	0	0	0	even parity being used
1	1	1	0	0	0	0	0	odd parity being used
0	1	0	1	0	1	0	1	even parity being used
1	1	1	0	0	0	1	1	odd parity being used

2 In each case, in **question 1**, where an error occurred, can you work out which bit in the byte was changed during data transmission?

If two of the bits change value following data transmission, it may be impossible to locate the error using parity checking.

Let us imagine we are transmitting the following byte, using even parity:

0	**1**	**0**	**1**	**1**	**1**	**0**	**0**

Suppose more than one bit has been modified during data transmission. This means the byte could have reached the destination as any of the following:

0	**1**	**1**	**1**	**1**	**1**	**0**	**1**	six 1-bits

0	1	0	1	0	0	0	0	two 1-bits

0	1	0	1	0	1	1	0	four 1-bits

In all these cases, the byte has clearly been corrupted, but the bytes have retained even parity. Therefore, no error would be flagged in spite of the obvious errors in transmission. Clearly it will be necessary to have other ways to complement parity when it comes to error checking to ensure errors are never missed. One such method is called checksum – see the next section.

You should have concluded that *any* of the bits in question 2 (Activity 2.5) could have been changed where there was a transmission error. Therefore, even though an error has been flagged, it is impossible to know **exactly** which bit is in error.

One of the ways round this problem is to use **parity blocks**. In this method, a block of data is sent and the number of 1-bits are totalled horizontally and vertically (in other words, a parity check is done in both horizontal and vertical directions). As the following example shows, this method not only identifies that an error has occurred but also indicates where the error is.

Example 1

In this example, nine bytes of data have been transmitted. Agreement has been made that **even parity** will be used. Another byte, known as the **parity byte**, has also been sent. This byte consists entirely of the parity bits produced by the vertical parity check. The parity byte also indicates the end of the block of data.

Table 2.3 shows how the data arrived at the receiving end. It is now necessary to check the parity of each byte horizontally (bytes 1 to 9) and vertically (columns 1 to 8). Each row and column where the parity has changed from even to odd should be flagged:

▼ **Table 2.3** Parity block showing nine bytes and parity byte

	Parity bit	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	Bit 8
Byte 1	1	1	1	1	0	1	1	0
Byte 2	1	0	0	1	0	1	0	1
Byte 3	0	1	1	1	1	1	1	0
Byte 4	1	0	0	0	0	0	1	0
Byte 5	0	1	1	0	1	0	0	1
Byte 6	1	0	0	0	1	0	0	0
Byte 7	1	0	1	0	1	1	1	1
Byte 8	0	0	0	1	1	0	1	0
Byte 9	0	0	0	1	0	0	1	0
Parity byte	1	1	0	1	0	0	0	1

A careful study of Table 2.3 shows the following:

- byte 8 (row 8) now has incorrect parity (there are three 1-bits)
- bit 5 (column 5) also now has incorrect parity (there are five 1-bits).

First of all, the table shows that an error has occurred following data transmission (there has been a change in parity in one of the bytes).

Secondly, at the intersection of row 8 and column 5, the position of the incorrect bit value (which caused the error) can be found. The 1-bit at this intersection should be a 0-bit; this means that byte 8 should have been:

0	0	0	1	0	0	1	0

which would also correct column 5 giving an even vertical parity (now has four 1-bits).

This byte could therefore be corrected automatically as shown above, or an error message could be relayed back to the sender asking them to re-transmit the block of data.

Activity 2.6

1 The following block of data was received after transmission from a remote computer; **odd parity** was being used by both sender and receiver. One of the bits has been changed during the transmission stage. Locate where this error is and suggest a corrected byte value:

	Parity bit	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	Bit 8
Byte 1	0	1	1	0	0	0	1	0
Byte 2	1	0	1	1	1	1	1	1
Byte 3	1	0	0	1	1	0	0	0
Byte 4	0	1	1	0	1	0	1	0
Byte 5	1	1	1	0	0	1	1	0
Byte 6	1	0	0	0	0	1	0	1
Byte 7	0	1	1	1	0	0	0	0
Byte 8	0	0	0	0	0	0	0	1
Byte 9	0	1	1	1	1	0	1	0
Parity byte	1	0	1	1	1	1	0	0

2 The following block of data was received after transmission from a remote computer. Even parity was being used by both sender and receiver. One of the bytes has been changed during the transmission stage. Locate where this error is and suggest a corrected byte value.

	Parity bit	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	Bit 8
Byte 1	1	1	0	0	0	0	0	0
Byte 2	0	0	1	1	1	1	0	0
Byte 3	0	1	0	0	0	1	1	1
Byte 4	1	0	1	0	1	1	1	1
Byte 5	0	0	0	1	0	0	0	1
Byte 6	0	0	1	1	1	1	1	1
Byte 7	1	0	1	1	0	1	0	0
Byte 8	0	1	0	1	0	0	0	1
Byte 9	1	1	1	0	0	1	0	0
Parity byte	0	0	1	1	1	0	0	1

Checksum

A **checksum** is a method used to check if data has been changed or corrupted following data transmission. Data is sent in blocks, and an additional value, called the checksum, is sent at the end of the block of data.

The checksum process is as follows:

- when a block of data is about to be transmitted, the checksum is calculated from the block of data
- the calculation is done using an agreed algorithm (this algorithm has been agreed by sender and receiver)
- the checksum is then transmitted with the block of data
- at the receiving end, the checksum is recalculated by the computer using the block of data (the agreed algorithm is used to find the checksum)
- the re-calculated checksum is then compared to the checksum sent with the data block
- if the two checksums are the same, then no transmission errors have occurred; otherwise a request is made to re-send the block of data.

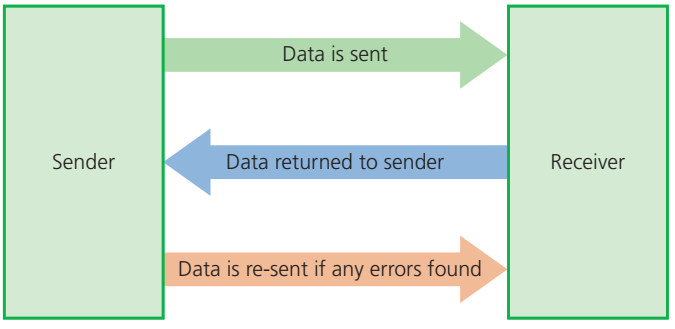
Echo check

With **echo check**, when data is sent to another device, this data is sent back again to the sender. The sender's computer compares the two sets of data to check if any errors occurred during the transmission process.

As you will have no doubt worked out, this isn't very reliable. If the two sets of data are different, it isn't known whether the error occurred when sending the data in the first place, or if the error occurred when sending the data back for checking.

However, if no errors occurred, then it is another way to check that the data was transmitted correctly. In summary:

- a copy of the data is sent back to the sender
- the returned data is compared with the original data by the sender's computer
- if there are no differences, then the data was sent without error
- if the two sets of data are different, then an error occurred at some stage during the data transmission.



▲ **Figure 2.14** Echo check diagram

2.2.3 Check digits

A **check digit** is the final digit included in a code; it is calculated from all the other digits in the code. Check digits are used for barcodes on products, such as International Standard Book Numbers (ISBN) and Vehicle Identification Numbers

(VIN). Check digits are used to identify errors in *data entry* caused by mis-typing or mis-scanning a barcode. They can usually detect the following types of error:

- an incorrect digit entered, for example 5327 entered instead of 5307
- transposition errors where two numbers have changed order, for example 5037 instead of 5307
- omitted or extra digits, for example 537 instead of 5307 or 53107 instead of 5307
- phonetic errors, for example 13 (thirteen), instead of 30 (thirty).

▲ **Figure 2.15** Sample barcode (ISBN 13 code with check digit)

There are a number of different methods used to generate a check digit. Two common methods will be considered here:

- ISBN 13
- Modulo-11

Example 1: ISBN 13

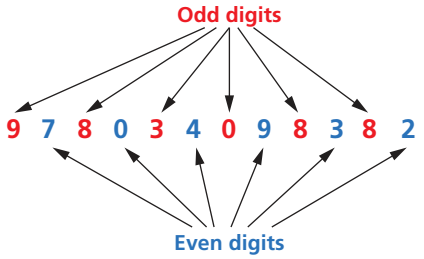
The check digit in ISBN 13 is the thirteenth digit in the number. We will now consider two different calculations. The first calculation is the generation of the check digit. The second calculation is a verification of the check digit (that is, a recalculation).

Calculation 1 – Generation of the check digit from the other 12 digits in a number

The following algorithm generates the check digit from the 12 other digits:

1. add all the odd numbered digits together
2. add all the even numbered digits together and multiply the result by 3
3. add the results from 1 and 2 together and divide by 10
4. take the remainder, if it is zero then use this value, otherwise subtract the remainder from 10 to find the check digit.

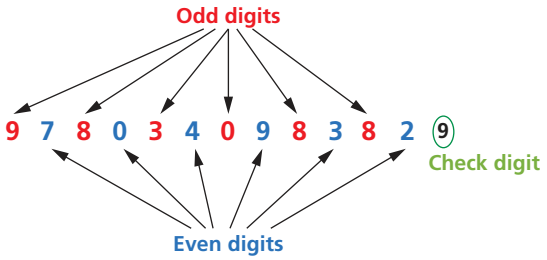
Using the ISBN **9 7 8 0 3 4 0 9 8 3 8 2** (note this is the same ISBN as in Figure 2.15):



▲ **Figure 2.16** ISBN (no check digit)

1. **9 + 8 + 3 + 0 + 8 + 8 = 36**
2. **3 × (7 + 0 + 4 + 9 + 3 + 2) = 75**
3. (**36** + **75**)/10 = 111/10 = 11 remainder 1
4. 10 – 1 = **9** the check digit

So we end up with the following thirteen-digit number (which matches the number shown in Figure 2.15):



▲ **Figure 2.17** ISBN (including the check digit)

Calculation 2 – Re-calculation of the check digit from the thirteen-digit number (which now includes the check digit)

To check that an ISBN 13-digit code is correct, including its check digit, a similar process is followed:

1 add all the odd numbered digits together, **including** the check digit
2 add all the even number of digits together and multiply the result by 3
3 add the results from 1 and 2 together and divide by 10
4 the number is correct if the remainder is zero.

Using the ISBN **9 7 8 0 3 4 0 9 8 3 8 2 9** (including its check digit) from Figure 2.17:

1 **9 + 8 + 3 + 0 + 8 + 8 + 9 = 45**
2 **3 × (7 + 0 + 4 + 9 + 3 + 2) = 75**
3 (45 + 75)/10 = 120/10 = 12 remainder **0**
4 remainder is 0, therefore number is correct.

Advice

You will not need to remember the steps shown in these algorithms; the steps will be given to you, but it is important that you understand how to use an algorithm to calculate or verify check digits.

? Example 2: Modulo-11

The modulo-11 method can have varying lengths of number which makes it suitable for many applications, such as product codes or VINs. The first calculation is the generation of the check digit. The second calculation is a verification of the check digit (that is, a recalculation).

Calculation 1 – Generation of the check digit from the other digits in a number

(In this example, we will assume the original number contained only 7 digits.)

The following algorithm generates the check digit from the other 7 digits:

1 each digit in the number is given a weighting of 8, 7, 6, 5, 4, 3 or 2 starting from the left (weightings start from 8 since the number will become eight-digit when the check digit is added)
2 the digit is multiplied by its weighting and then each value is added to make a total
3 the total is divided by 11
4 the remainder is then subtracted from 11 to find the check digit (note if the remainder is 10 then the check digit 'X' is used).

The example to be used has the following seven-digit number:

1 7-digit number: 4 1 5 6 7 1 0

weighting values: 8 7 6 5 4 3 2

2 sum: (8 × 4) + (7 × 1) + (6 × 5) + (5 × 6) + (4 × 7) + (3 × 1) + (2 × 0)
= 32 + 7 + 30 + 30 + 28 + 3 + 0
total = 130
3 divide total by 11: 130/11 = 11 remainder 9
4 subtract remainder from 11: 11 – 9 = 2 (check digit)

So we end up with the following eight-digit: 4 1 5 6 7 1 0 2

Calculation 2 – Re-calculation of the check digit from the eight-digit number (which now includes the check digit)

To check that the eight-digit number is correct, including its check digit, a similar process is followed:

1 each digit in the number is given a weighting of 8, 7, 6, 5, 4, 3, 2 or 1 starting from the left
2 the digit is multiplied by its weighting and then each value is added to make a total
3 the total is divided by 11
4 the number is correct if the remainder is zero

Using the 8-digit number: 4 1 5 6 7 1 0 2
1 weighting values: 8 7 6 5 4 3 2 1
2 sum: (8 × 4) + (7 × 1) + (6 × 5) + (5 × 6) + (4 × 7) + (3 × 1) + (2 × 0) + (1 × 2)
= 32 + 7 + 30 + 30 + 28 + 3 + 0 + 2
total = 132
3 divide total by 11: 132/11 = 12 remainder 0
4 remainder is 0, therefore number is correct

Activity 2.7

1 Using the algorithm for ISBN-13 calculate the check digit for:

978151045759

2 Find the check digits for the following numbers using **both** modulo-11 and ISBN 13 methods:
i 2 1 3 1 1 1 0 0 0 4 2 8
ii 9 0 9 8 1 2 1 2 3 5 4 4

2.2.4 Automatic Repeat Requests (ARQs)

We have already considered parity checks and echo checks as methods to verify that data has arrived at its destination unchanged. An **Automatic Repeat Request (ARQ)** is a third way used to check data following data transmission. This method can best be summarised as follows:

- ARQ uses positive and negative **acknowledgements** (messages sent to the receiver indicating that data has/has not been received correctly) and **timeout** (this is the time interval allowed to elapse before an acknowledgement is received)
- the receiving device receives an error detection code as part of the data transmission (this is typically a Cyclic Redundancy Check – refer to Section 2.1.1); this is used to detect whether the received data contains any transmission errors
- if no error is detected, a positive acknowledgement is sent back to the sending device
- however, if an error is detected, the receiving device now sends a negative acknowledgement to the sending device and requests re-transmission of the data
- a time-out is used by the sending device by waiting a pre-determined amount of time ….
- … and if no acknowledgement of any type has been received by the sending device within this time limit, it automatically re-sends the data until a positive acknowledgement is received ….
- … or until a pre-determined number of re-transmissions has taken place
- ARQ is often used by mobile phone networks to guarantee data integrity.

2.3 Symmetric and asymmetric encryption

2.3.1 The purpose of encryption

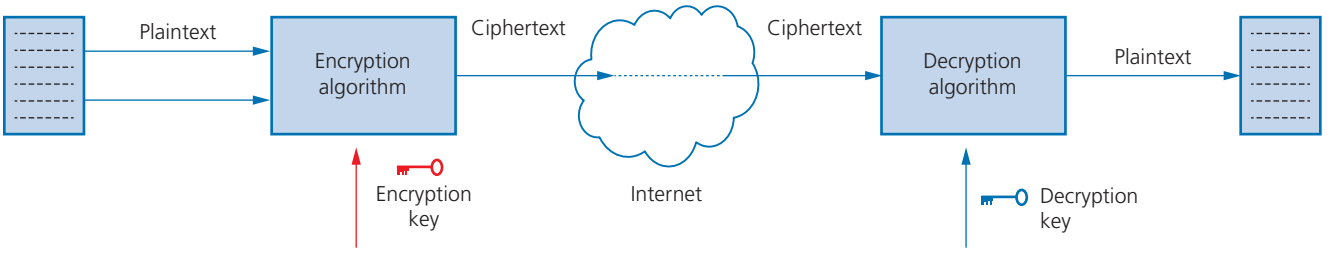
When data is transmitted over any public network (wired or wireless), there is always a risk of it being intercepted by, for example, a hacker. Under these circumstances, a hacker is often referred to as an **eavesdropper**. Using **encryption** helps to minimise this risk.

> **Link**
>
> For more on cyber security see Chapter 5.

Encryption alters data into a form that is unreadable by anybody for whom the data is not intended. It cannot prevent the data being intercepted, but it stops it from making any sense to the eavesdropper. This is particularly important if the data is sensitive or confidential (for example, credit card/bank details, medical history or legal documents).

Plaintext and ciphertext

The original data being sent is known as **plaintext**. Once it has gone through an **encryption algorithm**, it produces **ciphertext**:



▲ **Figure 2.18** Plaintext and ciphertext

2.3.2 Symmetric and asymmetric encryption

Symmetric encryption

Symmetric encryption uses an encryption key; the same key is used to encrypt and decrypt the encoded message. First of all, consider a simple system that uses a 10-digit denary encryption key (this gives 1×10^{10} possible codes); and a decryption key. Suppose our encryption key is:

4 2 9 1 3 6 2 8 5 6

which means every letter in a word is shifted across the alphabet +4, +2, +9, +1, and so on, places. For example, here is the message COMPUTER SCIENCE IS EXCITING (plaintext on the top line of Figure 2.19) before and after applying the encryption key (forming the ciphertext shown on the bottom line of Figure 2.19):

C	O	M	P	U	T	E	R
4	2	9	1	3	6	2	8
G	Q	V	Q	X	Z	G	Z

S	C	I	E	N	C	E
5	6	4	2	9	1	3
X	I	M	G	W	D	H

I	S
6	2
O	U

E	X	C	I	T	I	N	G
8	5	6	4	2	9	1	3
M	C	I	M	V	R	O	J

▲ **Figure 2.19** Plaintext into ciphertext using 10-digit encryption key

To get back to the original message, it will be necessary to apply the same decryption key; that is, 4 2 9 1 3 6 2 8 5 6. But in this case, the decryption process would be the reverse of encryption and each letter would be shifted –4, –2, –9, –1, and so on. For example, 'G' → 'C', 'Q' → 'O', 'V' → 'M', 'Q' → 'P', and so on.

However, modern computers could 'crack' this encryption key in a matter of seconds. To try to combat this, we now use 256-bit binary encryption keys that give 2^{256} (approximately, 1.2×10^{77}) possible combinations. (Even this may not be enough as we head towards **quantum computers**.)

Find out more

One of the ways of mitigating the risk of symmetric keys falling into the wrong hands (known as the key distribution problem) is to use a system based on modulo-11, where both sender and receiver can calculate the encryption key without it actually being exchanged in any way. Find out how this system works.

The real difficulty is keeping the encryption key a secret (for example, it needs to be sent in an email or a text message which can be intercepted). Therefore, the issue of security is always the main drawback of symmetrical encryption, since a single encryption key is required for both sender and recipient.

Asymmetric encryption

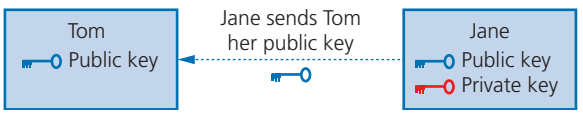
Asymmetric encryption was developed to overcome the security problems associated with symmetric encryption. It makes use of two keys called the **public key** and the **private key**:

- public key (made available to everybody)
- private key (only known to the computer user).

Both types of key are needed to encrypt and decrypt messages.

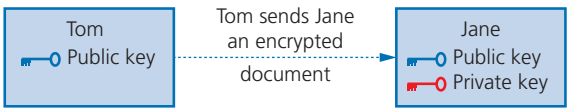
We will use an example to explain how this works; suppose Tom and Jane work for the same company and Tom wishes to send a confidential document to Jane:

1. Jane uses an algorithm to generate a **matching pair of keys** (private and public) that they must keep stored on their computers; the matching pairs of keys are mathematically linked but can't be derived from each other.
2. Jane now sends her public key to Tom.



▲ **Figure 2.20** Jane sends Tom her public key

3. Tom now uses Jane's public key (🔑) to encrypt the document he wishes to send to her. He then sends his encrypted document (ciphertext) back to Jane.



▲ **Figure 2.21** Encrypted document sent from Tom to Jane

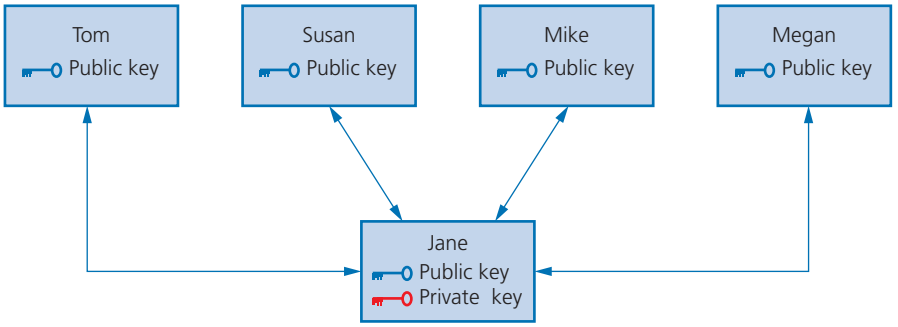
4. Jane uses her **matching** private key (🔑) to unlock Tom's document and decrypt it; this works because the public key used to encrypt the document and the private key used to decrypt it are a matching pair generated on Jane's computer. (Jane can't use the public key to decrypt the message.)

Activity 2.8

1 At the moment Jane can only receive encrypted documents from Tom.
Describe what would need to happen for Jane to be able to **send** encrypted documents back to Tom.

2 Explain why this method is much more secure than symmetric encryption.

Jane can also exchange her public key with any number of people working in the company, so she is able to receive encrypted messages (which have been encrypted using her public key) and she can then decrypt them using her matching private key:



▲ **Figure 2.22** The sharing of Jane's public key

However, if a two-way communication is required between all five workers, then they all need to generate their own matching public and private keys. Once this is done, all users then need to swap public keys so that they can send encrypted documents/files/messages between each other. Each worker will then use their own private key to decrypt information being sent to them.

Find out more

1 Using Figure 2.22 as your template, draw a new diagram showing the public keys and private keys that need to be swapped if Jane wishes to have a two-way exchange of encrypted documents between Tom, Susan, Mike and Megan.

2 Consider the complexity, if all five people want to have secure two-way communication between each other (and not just with Jane). This would mean each of the five workers sharing their own public keys with each of the other workers.

Activity 2.9

For each of the following ten questions, choose which of the five options corresponds to the correct response.

a What is meant by the term **ciphertext** when used in encryption?

A an encryption or decryption algorithm

B an encrypted message

C a type of session key

D another name for plaintext

E text following an encryption algorithm

b When carrying out asymmetric encryption, which of the following users would keep the private key?

A the sender
B the receiver
C both sender and receiver
D all recipients of the message
E none of the above

c In encryption, which of the following is the term used to describe the message before it is encrypted?

A simpletext
B plaintext
C notext
D ciphertext
E firsttext

d Which of the following is the biggest disadvantage of using symmetric encryption?

A it is very complex and time consuming
B it is rarely used anymore
C the value of the key reads the same in both directions
D it only works on computers with older operating systems
E there is a security problem when transmitting the encryption key

e Which of the following is the correct name for a form of encryption in which both the sender and the recipient use the same key to encrypt and decrypt?

A symmetric key encryption
B asymmetric key encryption
C public key encryption
D same key encryption
E block cipher encryption

f What of the following is the final number in a code, which is calculated from all the numbers in the code; its purpose is to find errors in data entry?

A parity check
B checksum
C cyclic redundancy check
D parity bit
E check digit

g Which of the following is a form of error detection that makes use of a system of acknowledgements and timeouts?

A automatic repeat request
B echo check
C check digit
D parity bit
E cyclic redundancy check

h Which of the following methods uses an extra bit added to a byte to ensure it contains an even number of 1s or odd number of 1s?

A cyclic redundancy check
B parity check
C checksum
D check digit
E echo check

i Which of the following uses a calculated value which is sent after a block of data; the receiving computer also calculates the value from the block of data and compares the values?

A parity check
B check digit
C packet switching
D checksum
E automatic repeat request

j Which of the following describes the check where the receiving computer sends back a copy of the data to the sending computer to allow it to compare the data?

A echo check
B automatic repeat request
C checksum
D parity check
E check digit

Extension

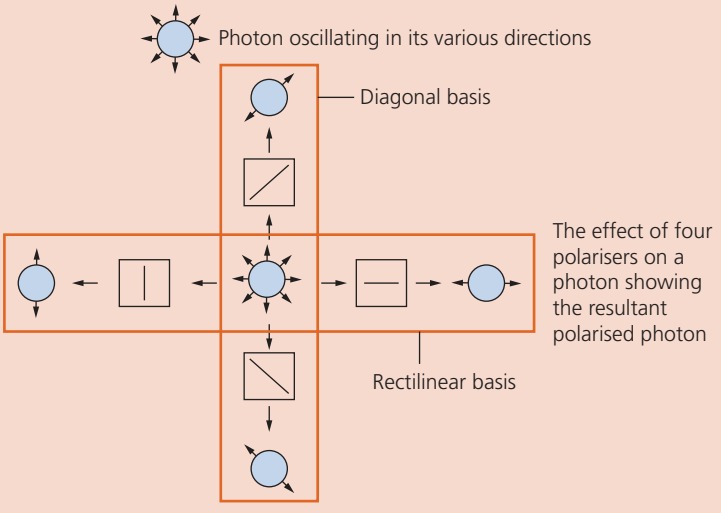
For those students considering the study of this subject at A Level, the following section gives some insight into further study on encryption.

Quantum cryptography exploits the laws of quantum mechanics to improve on the security of data. Quantum cryptography is based on the use of particles of light called photons (with energy calculated according to the formula: E = hf) and their physical quantum properties to produce a virtually unbreakable encryption system. This helps protect the security of data being transmitted over fibre optic cables. The technology is based on the laws of physics, rather than mathematics which is how the current cryptography methods already covered in this chapter work.

One of the uses of quantum cryptography is when sending encryption keys across a network – this uses a **Quantum Key Distribution (QKD)** protocol (one of the most common is called BB84).

QKD uses quantum mechanics to ensure a secure transmission of encryption keys. They use a **Qubit** (**Qu**antum **bit**) which is the basic unit of quantum 'data'. Unlike normal binary (which uses discrete 0s and 1s), the state of a Qubit is **both** 0 and 1 until the photon is measured.

A photon normally vibrates or oscillates in all different directions. Polarisation restricts these vibrations to particular directions. The following diagram shows how a photon can be polarised in one of two **bases** – the rectilinear (up/down and side-to-side) basis and the diagonal basis. Do not worry too much about what a basis is – you can just think of them as two different ways of preparing and measuring the photon.



▲ **Figure 2.23** Quantum cryptography

So how do we use quantum cryptography to send an encryption key from 'A' to 'B' using the QKD protocol? To understand this we need to note that:

- Photons can be polarised in one of two bases – rectilinear or diagonal.
- In the rectilinear basis, 1 can be represented by ↕ and 0 by ↔. In the diagonal basis 1 is ↗ and 0 is ↖.
- A photon that is polarised in the rectilinear basis will always give the same result when measured in the rectilinear basis. So, if a photon is polarised as '1' in the rectilinear basis (↕) it will always be measured in the rectilinear basis to have the polarisation ↕.
- Similarly, a photon that is polarised in the diagonal basis will always give the same result when measured in the diagonal basis. So, if a photon is polarised as '1' in the diagonal basis (↗) it will always be measured in the diagonal basis to have the polarisation ↗.
- However, if a photon is polarised in the rectilinear basis but measured in the diagonal basis then information about the original polarisation is lost. The result of the measurement has a 50-50 chance of being ↗ or ↖ because in the diagonal basis the photon is in both 1 and 0 states **at the same time!** This measurement tells the receiver nothing about the photon's original polarisation in the rectilinear basis.
- Similarly, if a photon is polarised in the diagonal basis but measured in the rectilinear basis then information about the original polarisation is lost. The result of the measurement has a 50-50 chance of being ↕ or ↔ because in the rectilinear basis the photon is in both 1 and 0 states **at the same time**! This tells the receiver nothing about the photon's original polarisation in the diagonal basis.

1. So, the sender polarises each photon using a basis that is selected at random.
2. The receiver measures each photon using a basis that is selected at random.
3. The sender and receiver publicly exchange which basis they used for each photon.

Only when the sender and receiver use the same basis for measurement can they be sure that are both reading a 1 or 0. When they used the same basis the receiver knows they have measured a '1' or '0' correctly.

In this chapter, you have learnt about:

- ✔ the use of data packets when sending data over networks
- ✔ packet switching
- ✔ types of data transmission (serial, parallel, simplex, half-duplex and full-duplex)
- ✔ the universal serial bus (USB)
- ✔ how errors during transmission can occur and how to recognise them and recover from them (using parity check, checksum and echo check)
- ✔ the use of check digits to identify errors following data entry
- ✔ how automatic repeat requests are used to find errors in transmitted data
- ✔ why encryption of data is used
- ✔ symmetric and asymmetric encryption
- ✔ the use of public keys and private keys in asymmetric encryption.

Key terms used throughout this chapter

data packet – a small part of a message/data that is transmitted over a network; after transmission all the data packets are reassembled to form the original message/data

packet header – the part of the data packet that contains the IP addresses of the sender and receiver, and includes the packet number which allows reassembly of the data packets

packet trailer – the part of a data packet that indicates the end of the data packet and cyclic redundancy check error check

cyclic redundancy check (CRC) – an error checking method in which all the 1-bits in the data packet payload are added and the total is stored in the packet trailer; the same calculation is repeated at the receiving station

payload – the actual data being carried in a data packet

node – stages in a network that can receive and transmit data packets; routers are nodes in communication networks

packet switching – a method of transmission in which a message is broken into many data packets which can then be sent along pathways independently of each other

router – a device that enables data packets to be moved between different networks, for example to join a LAN to a WAN

real time streaming – the transmission of data over a network for live events where the data is sent as soon as it is received or generated

hopping/hop number – a number in a data packet header used to stop data packets that never reach their destination from 'clogging up' the data paths/routes

simplex – data that can be sent on one direction only

half-duplex – data that can be sent in both directions but not at the same time

full-duplex – data that can be sent in both directions at the same time (simultaneously)

serial data transmission – sending data down one channel/wire one bit at a time

parallel data transmission – sending data down several channels/wires several bits at a time (usually 1 byte)

skewed (data) – data that arrives at the destination with the bits no longer synchronised

universal serial bus (USB) – a type of serial data transmission which has become the industry standard for connecting computers to devices via a USB port

parity check – a method used to check if data has been transferred correctly; it makes use of even parity (an even number of 1-bits) or odd parity (an odd number of 1-bits)

parity bit – a bit (either 0 or 1) added to a byte of data in the most significant bit position; this ensures that the byte follows the correct even parity or odd parity protocol

parity block – a horizontal and vertical parity check on a block of data being transmitted

parity byte – an extra byte of data sent at the end of a parity block; it is composed of the parity bits generated from a vertical parity check of the data block

checksum – a verification method used to check if data transferred has been altered or corrupted; calculated from the block of data of data being sent; the checksum value is sent after each data block

automatic repeat request (ARQ) – a method of checking transmitted data for errors; it makes use of acknowledgement and timeout to automatically request re-sending of data if the time interval before positive acknowledgement is too long

acknowledgement – a message sent to the receiver indicating that data has been received correctly (used in the ARQ error detection method)

timeout – the time interval allowed to elapse before an acknowledgement is received (in the ARQ error detection method)

echo check – a method used to check if data has been transferred correctly; data is sent to a receiver and then immediately sent back to the sender; the sender then checks if the received data matches the sent data

check digit – an additional digit appended to a number to check if the entered number is error-free; check digit is a data entry check and not a data transmission check

eavesdropper – another name for a hacker who intercepts data being transmitted on a wired or wireless network

encryption – the process of making data meaningless using encryption keys; without the correct decryption key the data cannot be decoded (unscrambled)

plaintext – the original text/message before it is put through an encryption algorithm

ciphertext – encrypted data that is the result of putting a plaintext message through an encryption algorithm

encryption algorithm – a complex piece of software that takes plaintext and generates an encrypted string known as ciphertext

symmetric encryption – a type of encryption in which the same encryption key is used both to encrypt and decrypt a message

asymmetric encryption – a type of encryption that uses public keys and private keys to ensure data is secure

public key – a type of encryption key that is known to all users

private key – a type of encryption key which is known only to the single computer/user

quantum computer – a computer that can perform very fast calculations; it can perform calculations that are based on probability rather than simple 0 or 1 values; this gives a quantum computer the potential to process considerably more data than existing computers

Exam-style questions

1 A company owns a number of vending machines. Data is sent from each of these machines at the end of the day. The data contains amount of money taken, products sold and any error conditions/reports.

a The company uses both echo checking and automatic repeat requests (ARQs).

i Describe how echo checks work. Explain whether this is a suitable error checking method in this application. [2]

ii Describe how automatic repeat request (ARQ) works. [3]

b ***Checksum*** and ***check digit*** are two terms often confused by students.

Describe **three** differences of the two techniques. [3]

2 Explain each of the following computer terms:

i packet switching

ii cyclic redundancy check

iii data skewing

iv universal serial bus

v parity bit. [10]

3 Eight descriptions are given in the following table. The table columns are labelled checksum, parity check and ARQ.
Tick (✓) the appropriate column which correctly matches each description to the error-checking technique. For each description, it is possible to match 1, 2, 3 or none of the error-checking methods. [8]

Description	Checksum ✓	Parity check ✓	ARQ ✓
extra bit sent with each byte of data			
makes use of timeout and acknowledgement			
if an error is found, a request is made to re-send the data			
check on whether a data packet has been changed following transmission			
re-calculation made on any additional data values sent to the recipient			
data is transmitted in blocks or packets			
a method that can determine which bit in a data stream has been changed			
additional value sent at the end of a block of data to be used to check if any data transmission errors occurred			

4 **a** Four statements about automatic repeat requests (ARQs) are given below, but they are not in the correct order. Put the statements into their correct sequence.

i the sending computer waits for a period of time to see if the receiving computer acknowledges receipt of the data

ii after a set time period, a timeout occurs which automatically triggers the re-sending of the data

iii the sending computer transmits a block of data to the receiving computer
iv this continues until the receiving computer sends an acknowledgement that the data has been received [3]

b Five statements about checksum error checking are given below, but they are not in the correct order. Put the statements into their correct sequence.

i if the two checksum values don't match, the receiving computer requests the data to be re-transmitted
ii the sending computer sends a block of data together with the checksum value
iii the receiving computer uses the block of data it receives to re-calculate the checksum using the same method as the sending computer
iv the two checksum values are compared by the receiving computer
v the sending computer uses the block of data to calculate the checksum using an agreed method [4]

c Five statements about parity checking are given below, but they are not in the correct order. Put the statements into their correct sequence.

i the sending computer sends the binary data including the parity bits
ii the sending and receiving computers agree the parity protocol (odd or even)
iii the sending computer adds a parity bit to each byte to make the byte odd or even parity
iv the receiving computer checks the parity of each byte received and checks it against the agreed protocol
v if the parity of the byte is incorrect, the receiving computer requests the data to be re-sent [4]

d Six statements about check digits are given below, but they are not in the correct order. Put the statements into their correct sequence.

i a human operator will be asked by the computer to re-enter the numerical code
ii the computer calculates the check digit based on the numerical code entered into the computer by a human operator
iii if the two check digits don't match, the human operator has made an error when entering the numerical code
iv the computer compares the calculated check digit with the check digit typed in by the human operator
v a human operator types in the numerical code into the computer
vi the check digit is calculated and added to the numerical code [5]

5 a Describe what is meant by ***symmetric encryption***. [2]

b Describe what is meant by ***asymmetric encryption***. [3]

c Explain why encryption is used when transmitting data over a network. [2]

6 Six descriptions are shown on the left and ten computer terms on the right.
By drawing lines, connect each description to the correct computer term (not all of the computer terms will be used). [6]

a method of error detection; a value is calculated from a block of data and is sent with the block of data during data transmission	skewed data
a method of error detection; it is based on counting the number of 1-bits; uses an additional bit which is the most significant bit in the byte	half-duplex
a data transmission method where data can be sent in both directions at the same time (simultaneously)	checksum
a data transmission method where data is sent one bit at a time over a single channel/wire	ARQ
a data error occurring when data arrives at the destination out of synchronisation	full-duplex
a form of serial data transmission which allows devices to communicate with a computer; it has become the industrial standard	check digit
	universal serial bus
	encryption
	serial
	parity check

7 A file server is used as a central data store for a network of computers. Rory sends data from his computer to a file server that is approximately 100 metres away. It is important that the data is transmitted accurately. Rory needs to be able to read data from and write data to the file server at the same time.

a i Use ticks (✓) to identify the most suitable data transmission methods for this application. [2]

Method 1	Tick ✓	Method 2	Tick ✓
Serial		Simplex	
Parallel		Half-duplex	
		Duplex	

ii Explain why your answer to **a i** is the most suitable data transmission. [4]

b Identify and describe **two** methods of error checking that can be used to make sure that the data stored after transmission is accurate. [6]

Cambridge O Level Computer Science 2210, Paper 12 Q4, Oct/Nov 2017

8 Maisey purchases a new router and attaches it to her computer.
The connection she sets up uses duplex data transmission.

a **Five** statements are given about duplex data transmission.
Tick (✓) to show if the statement is **True** or **False**. [5]

Statement	True ✓	False ✓
Duplex data transmission can be either serial or parallel		
Duplex data transmission is when data is transmitted both ways, but only one way at a time		
Duplex transmission is always used to connect a device to a computer		
Duplex data transmission is when data is transmitted both ways at the same time		
Duplex data transmission automatically detects any errors in data		

b Maisey's computer uses an integrated circuit (IC) for data transmission that sends multiple bits at the same time.
State whether the IC uses **serial** or **parallel** data transmission. [1]

c Maisey purchases a new printer and connects it to her computer using the USB port.
Explain **two** benefits of using a USB connection. [4]

Cambridge O Level Computer Science 2210, Paper 12 Q9, Oct/Nov 2019

3 Hardware

In this chapter you will learn about:

- ★ computer architecture
  - the Central Processing Unit (CPU)/microprocessor
  - von Neumann architecture
  - arithmetic and logic unit (ALU), control unit (CU) and registers
  - control bus, address bus, data bus
  - cores, cache and the internal clock
  - Fetch–Decode–Execute cycle
  - instruction set for a CPU
  - embedded systems
- ★ input and output devices
  - the following input devices: barcode and QR code scanners, digital cameras, keyboards, microphones, mouse, 2D/3D scanners and touch screens
  - the following output devices: actuators, light projectors, inkjet and laser printers, LED and LCD screens, speakers and 3D printers
  - sensors and their use in control and monitoring
- ★ data storage
  - primary storage (RAM and ROM)
  - secondary storage (magnetic, optical and solid state)
  - virtual memories
  - cloud storage
- ★ network hardware
  - network systems (NIC, MAC address, IP address and routers).

This chapter considers the hardware found in many computer systems. The hardware that makes up the computer itself and the various input and output devices will all be covered.

3.1 Computer architecture

3.1.1 The central processing unit (CPU)

The **central processing unit (CPU)** (also known as a microprocessor or processor) is central to all modern computer systems (including tablets and smartphones). The CPU is very often installed as an **integrated circuit** on a single microchip. The CPU has the responsibility for the execution or processing of all the instructions and data in a computer application. As Figure 3.1 shows, the CPU consists of:

- control unit (CU)
- arithmetic and logic unit (ALU)
- registers and buses.

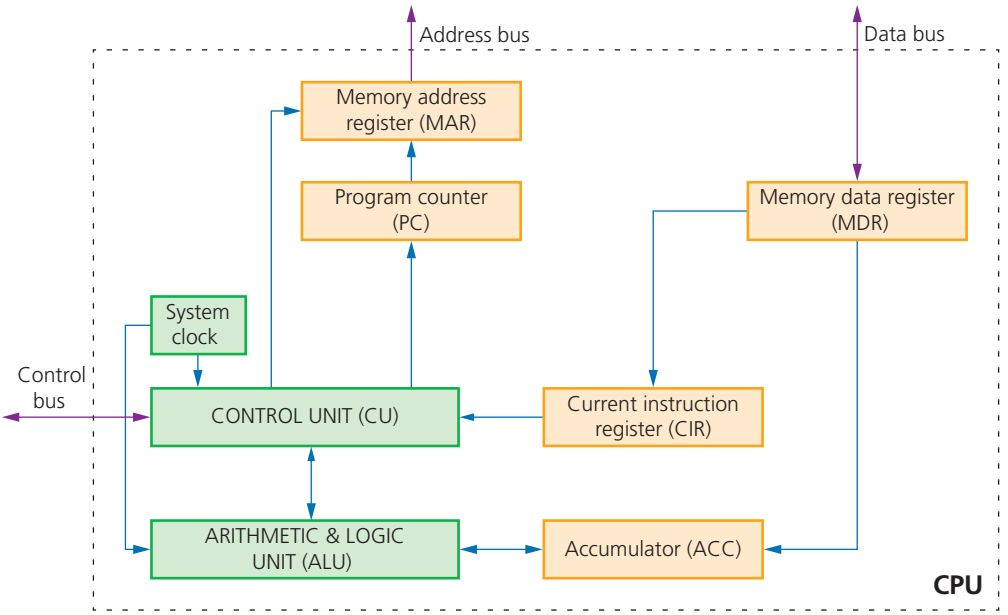
3.1.2 Von Neumann architecture

Early computers were fed data while the machines were actually running; it wasn't possible to store programs or data, which meant they couldn't operate without

considerable human intervention. In the mid-1940s, John von Neumann developed the concept of the 'stored program computer', which has been the basis of computer architecture ever since. The von Neumann architecture had the following main novel features (none of which were available in computers prior to the mid-1940s):

- the concept of a central processing unit (CPU or processor)
- the CPU was able to access the memory directly
- computer memories could store programs as well as data
- stored programs were made up of instructions which could be executed in sequential order.

There are many diagrams of von Neumann CPU architecture in other textbooks and on the internet. The following diagram is one example of a simple representation of von Neumann architecture:



▲ **Figure 3.1** Von Neumann architecture

Components of the central processing unit (CPU)

The main components of the CPU are the Control Unit (CU), Arithmetic & Logic Unit (ALU) and system clock.

Arithmetic & Logic Unit (ALU)

The **Arithmetic & Logic Unit (ALU)** allows the required arithmetic (e.g. +, - and shifting) or logic (e.g. AND, OR) operations to be carried out while a program is being run; it is possible for a computer to have more than one ALU to carry out specific functions. Multiplication and division are carried out by a sequence of addition, subtraction and left or right logical shift operations.

Link

For more arithmetic operations, please refer to Chapter 1.

Control Unit (CU)

The **control unit** reads an instruction from memory. The address of the location where the instruction can be found is stored in the Program Counter (PC). This instruction is then interpreted using the Fetch–Decode–Execute cycle (see later in this section). During that process, signals are generated along the control bus to tell the other components in the computer what to do. The control unit ensures synchronisation of data flow and program instructions throughout the computer. A

Link

For more details of the system clock, see Section 3.1.3.

Link

For more details on RAM, see Section 3.3.

system clock is used to produce timing signals on the control bus to ensure this vital synchronisation takes place – without the clock the computer would simply crash!

The RAM holds all the data and programs needed to be accessed by the CPU. The RAM is often referred to as the **Immediate Access Store (IAS)**. The CPU takes data and programs held in **backing store** (e.g. a hard disk drive) and puts them into RAM temporarily. This is done because read/write operations carried out using the RAM are considerably faster than read/write operations to backing store; consequently, any key data needed by an application will be stored temporarily in RAM to considerably speed up operations.

Registers

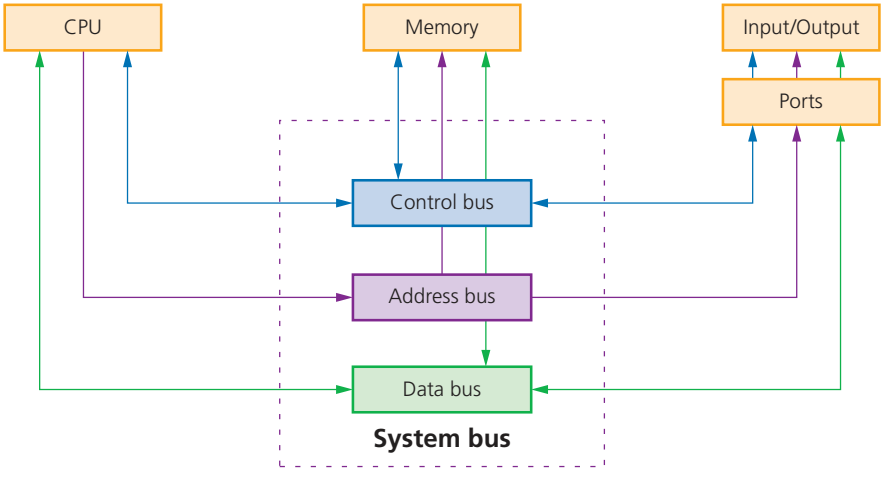
One of the most fundamental components of the von Neumann system are the **registers**. Registers can be general or special purpose. We will only consider the special purpose registers. A full list of the registers used in this textbook are summarised in Table 3.1. The use of these registers is explained more fully in the Fetch–Decode–Execute cycle (see later in this section).

▼ **Table 3.1** Specific purpose registers

Register	Abbreviation used	Function/purpose of register
current instruction register	CIR	this register stores the current instruction being decoded and executed
accumulator	ACC	this register is used when carrying out ALU calculations; it stores data temporarily during the calculations
memory address register	MAR	this register stores the address of the memory location currently being read from or written to
memory data/ buffer register	MDR	this register stores data which has just been read from memory or data which is about to be written to memory
program counter	PC	this register stores the address where the next instruction to be read can be found

System buses and memory

Earlier on, Figure 3.1 referred to some components labelled as buses. Figure 3.2 shows how these buses are used to connect the CPU to the memory and to input/output devices.



▲ **Figure 3.2** System buses and memory

▼ **Table 3.2** Section of computer memory

Address	Contents
1111 0000	0111 0010
1111 0001	0101 1011
1111 0010	1101 1101
1111 0011	0111 1011
1111 1100	1110 1010
1111 1101	1001 0101
1111 1110	1000 0010
1111 1111	0101 0101

Memory

The computer memory is made up of a number of partitions. Each partition consists of an **address** and its contents. Table 3.2 uses 8 bits for each address and 8 bits for the content. In a real computer memory, the address and its contents are actually much larger than this.
The address will uniquely identify every **location** in the memory and the contents will be the binary value stored in each location.

Let us now consider two examples of how the MAR and MDR registers can be used when carrying out a read and write operation to and from memory:

First, consider the **READ** operation. We will use the memory section shown in Table 3.2. Suppose we want to read the contents of memory location **1111 0001**; the two registers are used as follows:

- the address of location 1111 0001 to be read from is first written into the MAR (memory address register):

MAR:	1	1	1	1	0	0	0	1

- a 'read signal' is sent to the computer memory
- the contents of memory location 1111 0001 are then put into the MDR (memory data register):

MDR:	0	1	0	1	1	0	1	1

Now let us now consider the **WRITE** operation. Again, we will use the memory section shown in Table 3.2. Suppose this time we want to show how the value 1001 0101 was written into memory location 1111 1101:

- the data to be stored is first written into the MDR (memory data register):

MDR:	1	0	0	1	0	1	0	1

- this data has to be written into location with address: 1111 1101; so this address is now written into the MAR:

MAR:	1	1	1	1	1	1	0	1

- finally, a 'write signal' is sent to the computer memory and the value 10010101 will then be written into the correct memory location.

Input and output devices

The input and output devices will be covered in more detail in Section 3.2. They are the main method of entering data into and getting data out of computer systems. Input devices convert external data into a form the computer can understand and can then process (e.g. keyboards, touch screens and microphones). Output devices show the results of computer processing in a human understandable form (e.g. printers, monitors and loudspeakers).

(System) buses

(System) buses are used in computers as parallel transmission components; each wire in the bus transmits one bit of data. There are three common buses used in the von Neumann architecture known as: address bus, data bus and control bus.

Address bus

As the name suggests, the **address bus** carries addresses throughout the computer system. Between the CPU and memory, the address bus is **unidirectional** (i.e. bits can travel in one direction only); this prevents addresses being carried back to the CPU, which would be an undesirable feature.

The width of a bus is very important. The wider the bus, the more memory locations that can be directly addressed at any given time, e.g. a bus of width 16 bits can address 2^{16} (65 536) memory locations whereas a bus width of 32 bits allows 4 294 967 296 memory locations to be **simultaneously** addressed. However, even this isn't large enough for modern computers but the technology behind even wider buses is outside the scope of this book.

Data bus

The **data bus** is **bidirectional** (allowing data to be sent in both directions along the bus). This means data can be carried from CPU to memory (and vice versa) and to and from input/output devices. It is important to point out that data can be an address, an instruction or a numerical value. As with the address bus, the width of the data bus is important; the wider the bus the larger the **word length** that can be transported. (A **word** is a group of bits which can be regarded as a single unit e.g. 16-bit, 32-bit or 64-bit word lengths are the most common.) Larger word lengths can improve the computer's overall performance.

Control bus

The **control bus** is also bidirectional. It carries signals from the control unit (CU) to all the other computer components. It is usually 8-bits wide. There is no real need for it to be any wider since it only carries control signals.

Fetch–Decode–Execute cycle

To carry out a set of instructions, the CPU first of all **fetches** some data and instructions from memory and stores them in suitable registers. Both the address bus and data bus are used in this process. Once this is done, each instruction needs to be **decoded** before finally being **executed**. This is all known as the **Fetch–Decode–Execute cycle**.

Fetch

Both data and instruction can be stored in MDR. In the **Fetch–Decode–Execute cycle**, the next instruction is **fetched** from the memory address currently stored in the MAR and the instruction is stored in the MDR. The contents of the MDR are then copied to the Current Instruction Register (CIR). The PC is then incremented (increased by 1) so that the next instruction can be then be processed.

Decode

The instruction is then decoded so that it can be interpreted in the next part of the cycle.

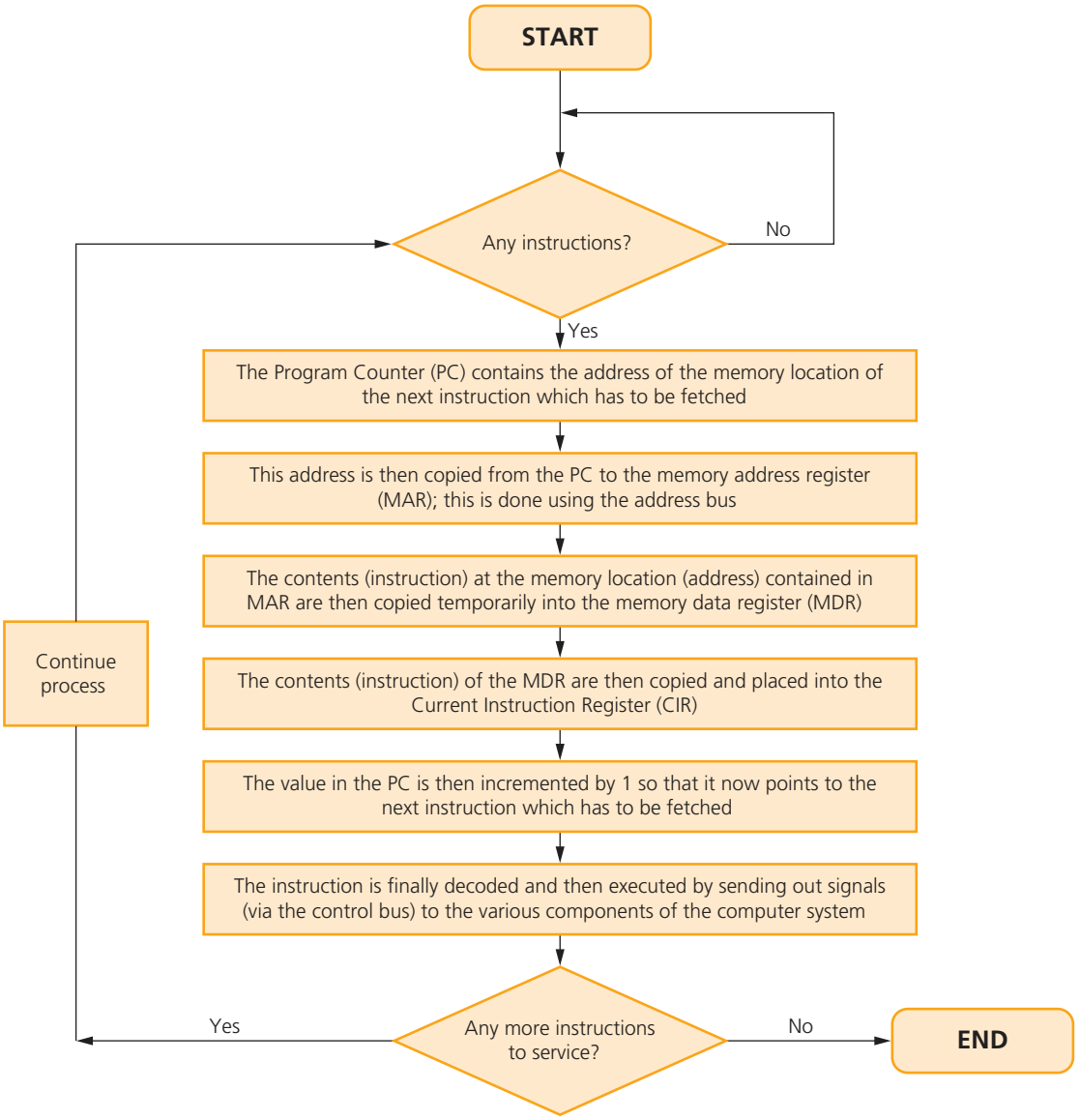
Execute

The CPU passes the decoded instruction as a set of control signals to the appropriate components within the computer system. This allows each instruction to be carried out in its logical sequence.

Link

See Section 4.2.2 for more details about these instructions.

Figure 3.3 shows how the Fetch–Decode–Execute cycle is carried out in the von Neumann computer model.



▲ **Figure 3.3** Fetch–Decode–Execute cycle flowchart

3.1.3 Cores, cache and internal clock

We will now consider the factors that determine the performance of a CPU. The first thing to consider is the role of the **system clock**. The clock defines the **clock cycle** that synchronises all computer operations. As mentioned earlier, the control bus transmits timing signals ensuring everything is fully synchronised. By increasing clock speed, the processing speed of the computer is also increased (a typical current value is 3.5 GHz – which means 3.5 billion clock cycles a second). Although the speed of the computer may have been increased, it isn't possible to say that a computer's overall *performance* is necessarily increased by using a higher clock speed. Other factors need to be considered, for example:

1 The width of the address bus and data bus (as mentioned earlier) can also affect computer performance and needs to be taken into account.

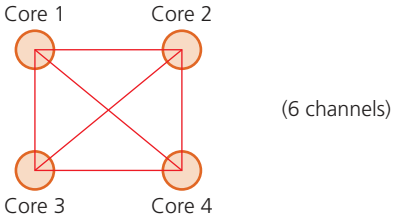
2 **Overclocking** is a factor to consider. The clock speed can be changed by accessing the **BIOS (Basic Input/Output System)** and altering the settings. However, using a clock speed higher than the computer was designed for can lead to problems, for example:
   - i execution of instructions outside design limits can lead to seriously unsynchronised operations (i.e. an instruction is unable to complete in time before the next one is due to be executed) – the computer would frequently crash and become unstable
   - ii overclocking can lead to serious overheating of the CPU again leading to unreliable performance.

3 The use of **cache** memories can also improve CPU performance. Unlike RAM, cache memory is located within the CPU itself, which means it has much faster data access times than RAM. Cache memory stores frequently used instructions and data that need to be accessed faster, which improves CPU performance. When a CPU wishes to read memory, it will first check out the cache and then move on to main memory/RAM if the required data isn't there. The larger the cache memory size the better the CPU performance.

4 The use of a different number of **cores** can improve computer performance. One core is made up of an ALU, a control unit and the registers. Many computers are dual core (the CPU is made up of two cores) or quad core (the CPU is made up of four cores). The idea of using more cores alleviates the need to continually increase clock speeds. However, doubling the number of cores doesn't necessarily double the computer's performance since we have to take into account the need for the CPU to communicate with each core; this will reduce overall performance.
   For example, with a **dual core** the CPU communicates with both cores using one channel reducing some of the potential increase in its performance:

▲ **Figure 3.4**

   while, with a **quad core** the CPU communicates with all four cores using six channels, considerably reducing potential performance:



▲ **Figure 3.5**

So all these factors need to be taken into account when considering computer performance. Summarising these points:

- increasing bus width (data and address buses) increases the performance and speed of a computer system
- increasing clock speed will potentially increase the speed of a computer

- a computer's performance can be changed by altering bus width, clock speed and use of multi-core CPUs
- use of cache memories can also speed up a CPU's performance.

Activity 3.1

1 a Name three buses used in the von Neumann architecture.

b Describe the function of each named bus.

c Describe how bus width and clock speed can affect computer performance.

2 Complete the following paragraph by using terms from this chapter:

The CPU ..... data and instructions required for an application and temporarily stores them in the ..... until they can be processed. The ..... is used to hold the address of the next instruction to be executed. This address is copied to the ..... using the ..... The contents at this address are stored in the ..... Each instruction is then ..... and finally ..... by sending out ..... using the ..... Any calculations carried out are done using the ..... During any calculations, data is temporarily held in a special purpose register known as the .....

3.1.4 Instruction set

In a computer system, instructions are a set of operations which are decoded in sequence. Each operation will instruct the ALU and CU (which are part of the CPU). An operation is made up of an **opcode** and an **operand**.

The opcode informs the CPU what operation needs to be done

The operand is the data which needs to be acted on or it can refer to a register in the memory

Since the computer needs to understand the operation to be carried out, there is actually a limited number of opcodes that can be used; this is known as the **instruction set**. All software running on a computer will contain a set of instructions (which need to be converted into binary). The Fetch–Decode–Execute cycle is the sequence of steps used by the CPU to process each instruction in sequence.

One example of an instruction set is the X86, a common CPU standard used in many modern computers. Although different computer manufacturers will adopt their own internal electronic design, if the computer is based on the X86 CPU then all designs will share almost identical instruction sets. For example, Intel Pentium and AMD Athlon CPUs use almost identical X86 instruction sets even though they are based on very different electronic designs.

(Note of caution: do not confuse instruction sets with programming code; instruction sets are the low-level language instructions that instruct the CPU how to carry out an operation. Program code needs interpreters or compilers to convert the code into the instruction set understood by the computer. Some examples of instruction set operations include: ADD, JMP, LDA, and so on.)

Link

See Section 4.2.2 for examples of opcodes and operands in assembly language.

Link

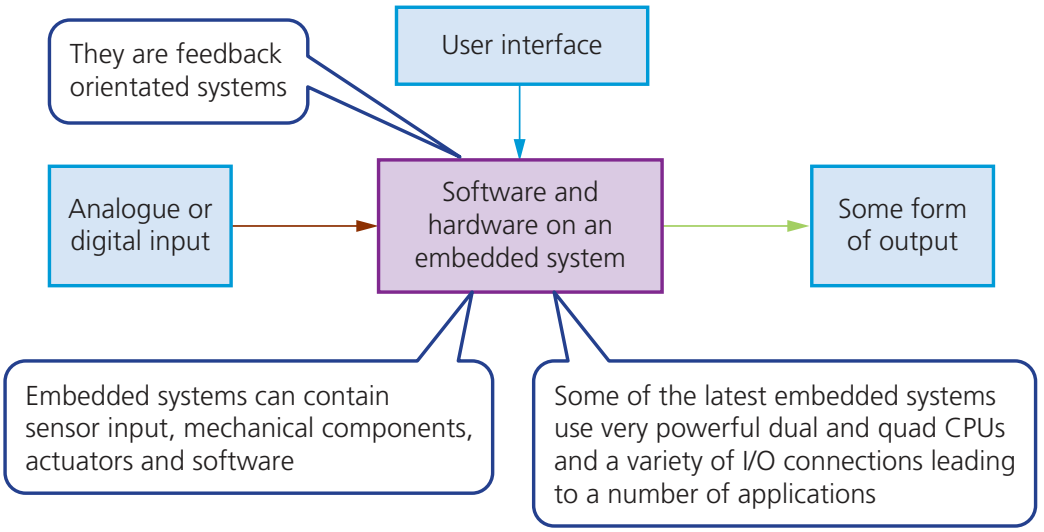
For more on interpreters and compilers see Section 4.2.3.

3.1.5 Embedded systems

An embedded system is a combination of hardware and software which is designed to carry out a specific set of functions. The hardware is electronic, electrical or electro-mechanical. Embedded systems can be based on:

microcontrollers:	this has a CPU in addition to some RAM and ROM and other peripherals all embedded onto one single chip (together they carry out a specific task)
microprocessor:	integrated circuit which only has a CPU on the chip (there is no RAM, ROM or peripherals – these need to be added)
system on chips (SoC):	this may contain a microcontroller as one of its components (they almost always will include CPU, memory, input/output (I/O) ports and secondary storage on a single microchip)

An embedded system will have a specific set of tasks; Figure 3.6 summarises how embedded systems work in general:



▲ **Figure 3.6** Embedded systems

When installed in a device, either an operator can input data manually (for example, select a temperature from a keypad or turn a dial on an oven control panel) or the data will come from an automatic source, such as a sensor. This sensor input will be analogue or digital in nature, for example, inputs such as oxygen levels or fuel pressure in a car's engine management system. The output will then carry out the function of the embedded system by sending signals to the components that are being controlled (for example, increase the power to the heating elements in an oven or reduce fuel levels in the engine).

Depending on the device, embedded systems are either programmable or non-programmable. Non-programmable devices need, in general, to be replaced if they require a software upgrade. Programmable devices permit upgrading by two methods:

- connecting the device to a computer and allowing the download of updates to the software (for example, this is used to update the maps on a GPS system used in a vehicle)

- automatic updates via a Wi-Fi, satellite or cellular (mobile phone network) link (for example, many modern cars allow updates to engine management systems and other components via satellite link).

There are definite benefits and drawbacks of devices being controlled using embedded systems:

Table 3.3 Benefits and drawbacks of using embedded systems

Benefits	Drawbacks
they are small in size and therefore easy to fit into devices	it can be difficult to upgrade some devices to take advantage of new technology
compared to other systems, they are relatively low cost to make	troubleshooting faults in the device becomes a specialist task
they are usually dedicated to one task allowing simple interfaces and often no requirement for an operating system	although the interface can appear to be more simple (e.g. a single knob) in reality it can be more confusing (e.g. changing the time on a cooker clock can require several steps!)
they consume very little power	any device that can be accessed over the internet is also open to hackers, viruses, etc.
they can be controlled remotely using a mobile phone, for example	due to the difficulty in upgrading and fault finding, devices are often just thrown away rather than being repaired (very wasteful)
very fast reaction to changing input (operate in real time and are feedback orientated)	can lead to an increase in the 'throw away' society if devices are discarded just because they have become out-of-date
with mass production comes reliability	

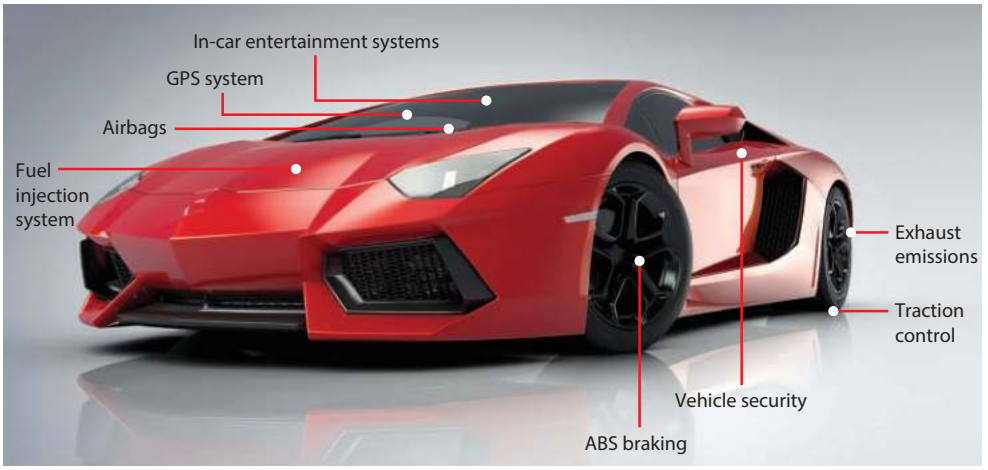
Because embedded systems can be connected to the internet, it is possible to control them remotely using a smartphone or computer. For example, setting the central heating system to switch on or off while away from home or remotely instructing a set top box to record a television programme. Since embedded systems are dedicated to a specific set of tasks, engineers can optimise their designs to reduce the physical size and cost of the devices. The range of applications are vast, ranging from a single microcontroller (for example, in an MP3 player) to a complex array of multiple units (for example, in a medical imaging system).

It is worth mentioning here that a computer is ***not*** an example of an embedded system. Computers are multi-functional (that is, they can carry out many different tasks which can be varied by using different software) which means they can't be classed as embedded systems.

Examples of the use of embedded systems

Motor vehicles

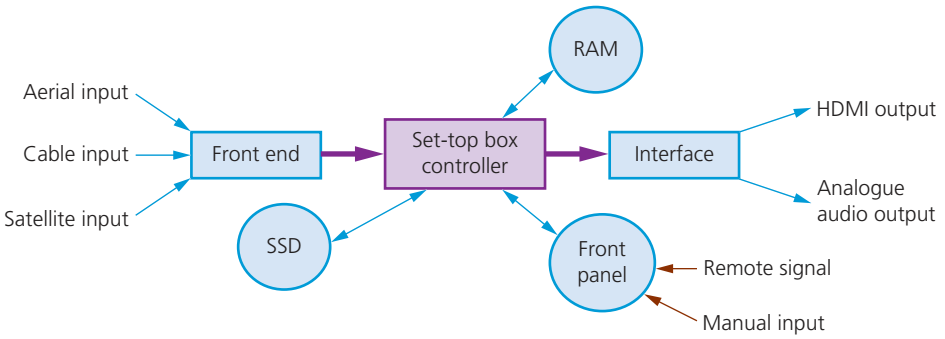
Modern cars have many parts that rely on embedded systems to function correctly. Figure 3.7 shows some of the many components that are controlled in this way.



▲ **Figure 3.7** Embedded systems found in a car

Set-top box

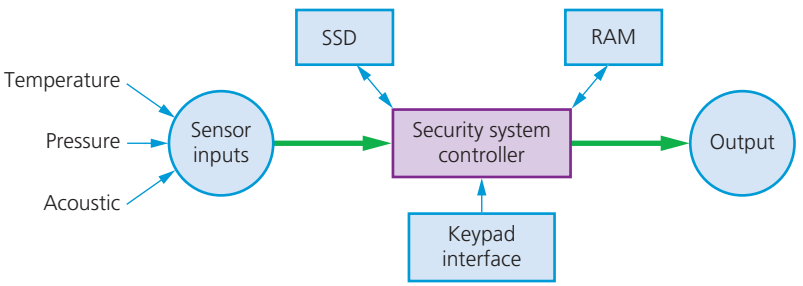
In this example, a set-top box uses an embedded system to allow, for example, recording and playback of television programmes. This can be operated remotely by the user when not at home using an internet-enabled device or by using the interface panel when at home. The embedded system will look after many of the functions involving inputs from a number of sources such as a Solid State Device (SSD) (where television programmes can be stored or retrieved) or a satellite signal (where it will be necessary to decode the incoming signal).



▲ **Figure 3.8** Embedded system found in a set-top box

Security systems

Embedded systems are used in many security devices:



▲ **Figure 3.9** Embedded system found in a security system

The security code is set in RAM and the alarm activated or deactivated using the keypad. Data from sensors is sent to the controller which checks against values stored on the SSD (these settings are on SSD rather than RAM in case

the sensitivity needs to be adjusted). An output can be a signal to flash lights, sound an alarm or send a message to the home owner via their mobile phone. Again, the home owner can interface with the system remotely if necessary.

Link

See Section 3.2.3 for sensors used in security systems.

Lighting systems

Embedded systems are used in modern sophisticated lighting systems from simple home use to major architectural lighting systems. We will concentrate here on a lighting system used in a large office. The system needs to control the lighting taking into account:

- the time of day or day of the week
- whether the room is occupied
- the brightness of the natural light.

An embedded system can automatically control the lighting using a number of inputs (such as light sensors) and key data stored in memory. Again, this fits very well into the system described in Figure 3.6.

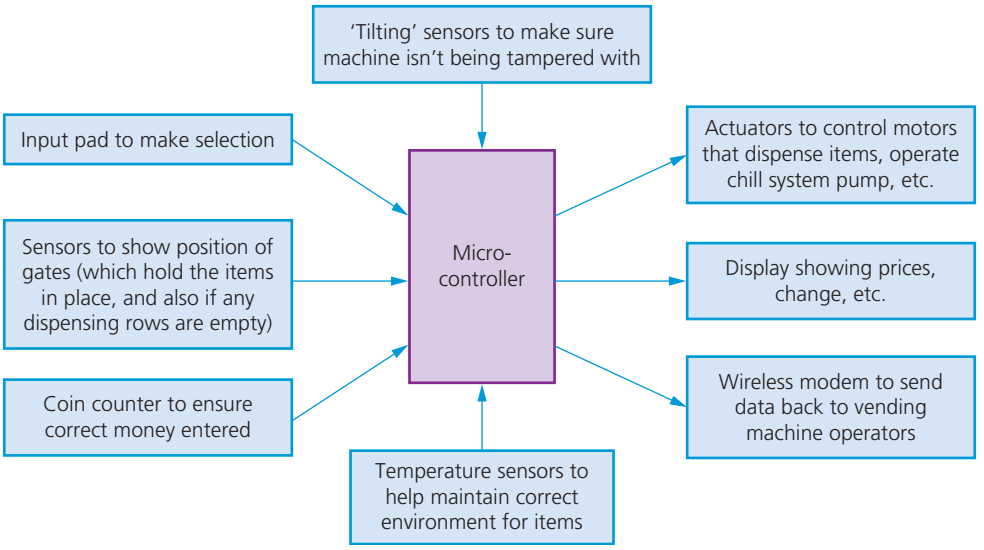
The time of day or day of the week is important data in an office environment since energy is saved if the system switches to low lighting levels when unoccupied. This can be over-ridden by the second bullet point (above); if there is movement in the office then correct lighting levels will be automatically restored. On a very bright sunny day, the system could automatically dim the lights, only increasing the light output if natural light levels fall below a set value. There are many internal and external lighting systems that could be controlled by embedded systems (e.g. a fountain light display or a light show on a building to commemorate a special occasion). They are also used to trigger emergency lighting in, for example, aeroplanes in case of an emergency.

Some lighting systems use Bluetooth light bulbs. This allows the embedded system to control each bulb independently. Many of the bulbs available today use LEDs and many come in a number of colours to change the mood.

▲ **Figure 3.10** LED light bulb

Vending systems

Vending machines make considerable use of embedded systems. They usually use microcontrollers to control a number of functions that we all associate with vending machines:



▲ **Figure 3.11** Embedded system found in a vending machine

At the heart of the vending machine is an embedded system in the form of a microcontroller. Inputs to this system come from the keypad (item selection) and from sensors (used to count the coins inserted by the customer, the temperature inside the machine and a 'tilt sensor' for security purposes). The outputs are:

- actuators to operate the motors, which drive the helixes (see figure below) to give the customers their selected item(s)
- signals to operate the cooling system if the temperature is too high
- item description and any change due shown on an LCD display panel
- data sent back to the vending machine company so that they can remotely check sales activity (which could include instructions to refill the machine) without the need to visit each machine.

▲ **Figure 3.12** Helix used in a typical vending machine

All of this is controlled by an embedded system which makes the whole operation automatic but also gives immediate sales analysis which would otherwise be very time consuming.

Washing machines

Many 'white goods' (such as refrigerators, washing machines, microwave ovens, and so on) are controlled by embedded systems. They all come with a keypad or dials that are used to select the temperature, wash cycle or cooking duration. This data forms the input to the embedded system, which then carries out the required task without any further human intervention. As with other devices, these 'white goods' can also be operated remotely using an internet-enabled smartphone or computer.

Activity 3.2

1 a Explain how it is possible to increase the performance of a CPU/microprocessor. In your explanation, include some of the risks associated with your suggestions to improve performance.

b What is meant by the term ***instruction set***?

2 A car is fitted with the latest GPS navigation system. This device is controlled by an embedded system in the form of a microcontroller.

a Describe the inputs needed by the embedded system and describe which outputs you would expect it to produce.

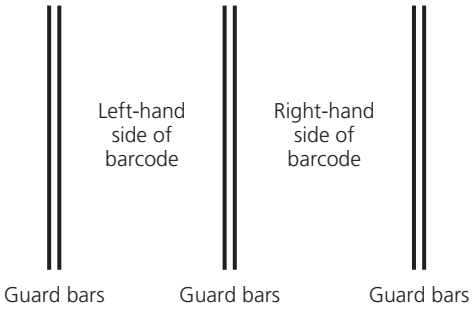
b Since updates to the GPS device are required every six months, explain how the device is updated without the need to take the car to the garage every six months.

3.2 Input and output devices

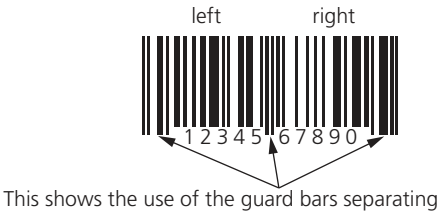
3.2.1 Input devices

Barcode scanners (readers)

A **barcode** is a series of dark and light parallel lines of varying thickness. The numbers 0 to 9 are each represented by a unique series of lines. Various barcode methods for representing these digits exist. The example we shall use adopts different codes for digits appearing on the left and for digits appearing on the right of the barcode:

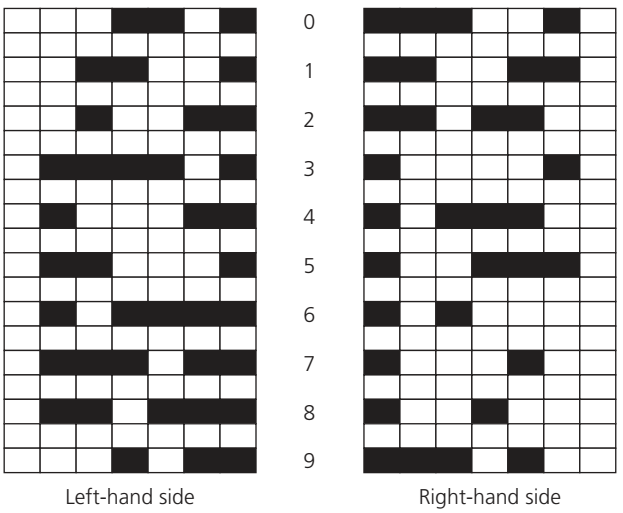


▲ **Figure 3.13** Diagram of guard bars



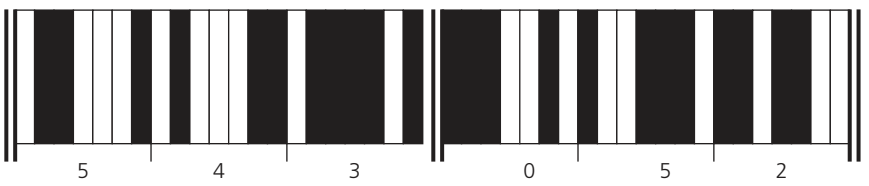
▲ **Figure 3.14** Sample barcode

Each digit in the barcode is represented by bars of 1 to 4 blocks thick as shown in Figure 3.15. Note there are different patterns for digits on the left-hand side and for digits on the right-hand side.



▲ **Figure 3.15** Barcode digit patterns

The section of barcode to represent the number 5 4 3 0 5 2 would therefore be:



▲ **Figure 3.16** Sample barcode section using patterns from Figure 3.15

Each digit is made up of 2 dark lines and two light lines. The width representing each digit is the same. The digits on the left have an odd number of dark elements and always begin with a light bar; the digits on the right have an even number of dark elements and always begin with a dark bar. This arrangement allows a barcode to be scanned in any direction.

So what happens when a barcode is scanned?

- the barcode is first of all read by a red laser or red LED (light emitting diode)
- light is reflected back off the barcode; the dark areas reflect little or no light, which allows the bars to be read
- the reflected light is read by sensors (photoelectric cells)
- as the laser or LED light is scanned across the barcode, a pattern is generated, which is converted into digital data – this allows the computer to understand the barcode
- for example: the digit '3' on the left generates the pattern: L D D D D L D (where L = light and D = dark),
  this has the binary equivalent of: 0 1 1 1 1 0 1 (where L = 0 and D = 1).

Barcodes are most commonly found at the checkout in supermarkets. There are several other input and output devices at the checkout:

▼ **Table 3.4** Input and output devices at a checkout

Input/output device	How it is used
keypad	to key in the number of same items bought; to key in a weight, to key in the number under the barcode if it cannot be read by the barcode reader/scanner
screen/monitor	to show the cost of an item and other information
speaker	to make a beeping sound every time a barcode is read correctly; but also to make another sound if there is an error when reading the barcode
printer	to print out a receipt/itemised list
card reader/chip and PIN	to read the customer's credit/debit card (either using PIN or contactless)
touchscreen	to select items by touching an icon (such as fresh fruit which may be sold loose without packaging)

So the barcode has been read, then what happens?

- the barcode number is looked up in the stock database (the barcode is known as the **key field** in the stock item record); this key field uniquely identifies each stock item
- when the barcode number is found, the stock item record is looked up
- the price and other stock item details are sent back to the checkout (or point of sale terminal (POS))
- the number of stock items in the record is reduced by 1 each time the barcode is read
- this new value for number of stock is written back to the stock item record
- the number of stock items is compared to the re-order level; if it is less than or equal to this value, more stock items are ***automatically*** ordered

- once an order for more stock items is generated, a flag is added to the record to stop re-ordering every time the stock item barcode is read
- when new stock items arrive, the stock levels are updated in the database.

Advantages to the management of using barcodes

- much easier and faster to change prices on stock items
- much better, more up-to-date sales information/sales trends
- no need to price every stock item on the shelves (this reduces time and cost to the management)
- allows for automatic stock control
- possible to check customer buying habits more easily by linking barcodes to, for example, customer loyalty cards.

Advantages to the customers of using barcodes

- faster checkout queues (staff don't need to remember/look up prices of items)
- errors in charging customers is reduced
- the customer is given an itemised bill
- cost savings can be passed on to the customer
- better track of 'sell by dates' so food should be fresher.

The barcode system is used in many other areas. For example, barcodes can be utilised in libraries where they are used in books and on the borrower's library card. Every time a book is taken out, the borrower is linked to the book automatically. This allows automatic checking of when the book is due to be returned.

▲ **Figure 3.17** Sample QR code

Quick response (QR) codes

Another type of barcode is the **quick response (QR) code**. This is made up of a matrix of filled-in dark squares on a light background. For example, the QR code in Figure 3.17 is a website advertising rock music merchandise. It includes a web address in the code.

QR codes can hold considerably more information than the more conventional barcodes described earlier.

Description of QR codes

- A QR code consists of a block of small squares (light and dark) known as pixels. It can presently hold up to 4296 characters (or up to 7089 digits) and also allows internet addresses to be encoded within the QR code. This compares to the 30 digits that is the maximum for a barcode. However, as more and more data is added, the structure of the QR code becomes more complex.
- The three large squares at the corners of the code function as a form of alignment; the remaining small corner square is used to ensure the correct size and correct angle of the camera shot when the QR code is read.

Because of modern smartphones and tablets, which allow internet access on the move, QR codes can be scanned anywhere. This gives rise to a number of uses:

- advertising products (for example, the QR code in Figure 3.17)
- giving automatic access to a website or contact telephone number
- storing boarding passes electronically at airports and train stations (Figure 3.18).

▲ **Figure 3.18** Sample boarding pass

By using the built-in camera on a mobile smartphone or tablet and by downloading a QR app (application), it is possible to read QR codes on the move using the following method:

- point the phone or tablet camera at the QR code
- the app will now process the image taken by the camera, converting the squares into readable data
- the browser software on the mobile phone or tablet automatically reads the data generated by the app; it will also decode any web addresses contained within the QR code
- the user will then be sent to a website automatically (or if a telephone number was embedded in the code, the user will be sent to the phone app)
- if the QR code contained a boarding pass, this will be automatically sent to the phone/tablet.

Advantages of QR codes compared to traditional barcodes

- They can hold much more information
- There will be fewer errors; the higher capacity of the QR code allows the use of built-in error-checking systems – normal barcodes contain almost no data redundancy (data which is duplicated) therefore it isn't possible to guard against badly printed or damaged barcodes
- QR codes are easier to read; they don't need expensive laser or LED (light emitting diode) scanners like barcodes – they can be read by the cameras on smartphones or tablets
- It is easy to transmit QR codes either as text messages or images
- It is also possible to encrypt QR codes which gives them greater protection than traditional barcodes.

Disadvantages of QR codes compared to traditional barcodes

- More than one QR format is available
- QR codes can be used to transmit malicious codes – known as attagging. Since there are a large number of free apps available to a user for generating QR codes, that means anyone can do this. It is relatively easy to write malicious code and embed this within the QR code. When the code is scanned, it is possible the creator of the malicious code could gain access to everything on the user's phone (for example, photographs, address book, stored passwords, and so on). The user could also be sent to a fake website or it is even possible for a virus to be downloaded.

▲ **Figure 3.19** Frame QR code

New developments

Newer QR codes (called **frame QR codes**) are now being used because of the increased ability to add advertising logos (see Figure 3.19). Frame QR codes come with a 'canvas area' where it is possible to include graphics or images inside the code itself. Unlike normal QR codes, software to do this isn't usually free.

Activity 3.3

1 Using the data in Figure 3.14, design the barcodes for:
   - **a** 900 340 (3 digits on the left; 3 digits on the right)
   - **b** 1257 6648 (4 digits on the left; 4 digits on the right)
   - **c** 05889 02918 (5 digits on the left; 5 digits on the right)

2 **a** Describe one advantage of using QR codes rather than traditional bar codes. Explain how barcodes bring the advantage you have described.
   - **b** A square QR code contains 40 × 40 tiny squares (pixels) where each tiny square represents a 0 or a 1. Calculate how many bytes of data can be stored on the QR code.
   - **c** Describe the purpose of the three large squares at the corners of the QR code.
   - **d** Describe one disadvantage of using QR codes.

▲ **Figure 3.20** Digital camera

Digital cameras

Digital cameras have essentially replaced the more traditional camera that used film to capture the images. The film required developing and then printing before the photographer could see the result of their work.

This made these cameras expensive to operate since it wasn't possible to delete unwanted photographs.

Modern digital cameras simply link to a computer system via a USB port or by using Bluetooth (which enables wireless transfer of photographic files).

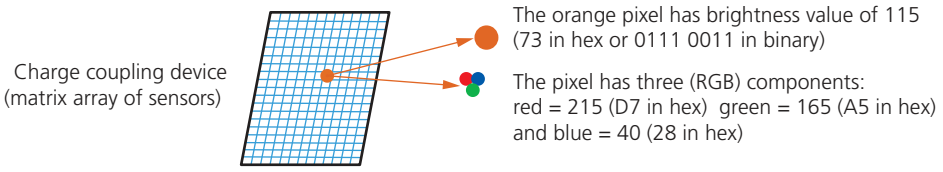
These cameras are controlled by an embedded system which can automatically carry out the following tasks:

- adjust the shutter speed
- focus the image automatically
- operate the flash gun automatically
- adjust the aperture size
- adjust the size of the image
- remove 'red eye' when the flash gun has been used
- and so on.

What happens when a photograph is taken

- the image is captured when light passes through the lens onto a light-sensitive cell; this cell is made up of millions of tiny sensors which are acting as photodiodes (i.e. **charge couple devices (CCD)** which convert light into electricity)
- each of the sensors are often referred to as pixels (picture elements) since they are tiny components that make up the image
- the image is converted into tiny electric charges which are then passed through an **analogue to digital converter (ADC)** to form a digital image array
- the ADC converts the electric charges from each pixel into levels of brightness (now in a digital format); for example, an 8-bit ADC gives 2^8 (256) possible brightness levels per pixel (for example, brightness level 01110011)

- apart from brightness, the sensors also measure colour which produces another binary pattern; most cameras use a 24-bit RGB system (each pixel has 8 bits representing each of the 3 primary colours), which means each pixel has a red value (0 to 255 in denary), a green value (0 to 255) and a blue value (0 to 255); for example, a shade of orange could be 215 (red), 165 (green) and 40 (blue) giving a binary pattern of 1101 0111 1010 0101 0010 1000 (or D7A528 written in hex)



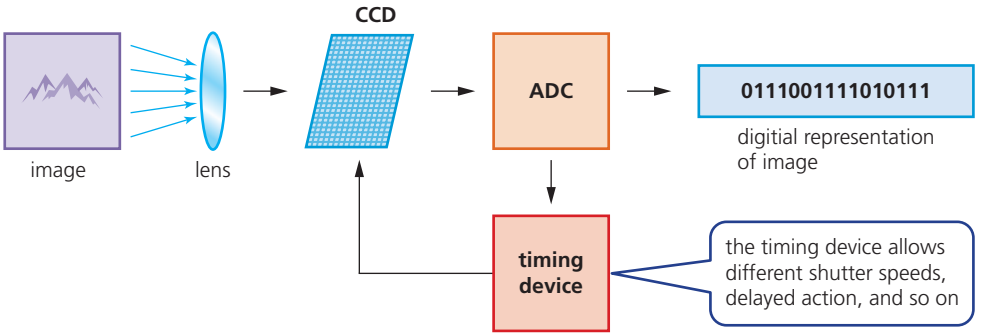
▲ **Figure 3.21** Typical pixel brightness and colour values

- the number of pixels determines the size of the file used to store the photograph
- the quality of the image depends on the recording device (how good the camera lens is and how good the sensor array is), the number of pixels used (the more pixels used, the better the image), the levels of light and how the image is stored (JPEG, raw file, and so on).

Link

For an explanation of how pixels affect file size see Chapter 1.

Mobile phones have caught up with digital cameras as regards number of pixels. But the drawback is often inferior lens quality and limited memory for the storage of photos. But this is fast changing and, at the time of writing, many smartphones now have very sophisticated optics and photography software as standard.



▲ **Figure 3.22** Diagram of how a digital camera works

Keyboards

Keyboards are by far the most common method used for data entry. They are used as the input devices on computers, tablets, mobile phones and many other electronic items.

The keyboard is connected to the computer either by using a USB connection or by wireless connection. In the case of tablets and mobile phones, the keyboard is often **virtual** or a type of **touch screen** technology.

▲ **Figure 3.23** Keyboard

▲ **Figure 3.24** Ergonomic keyboard

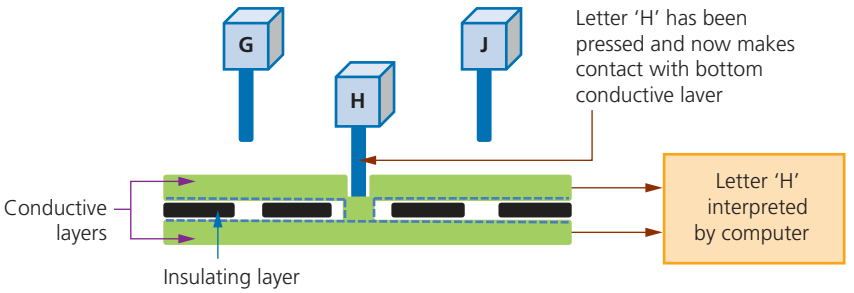
As shown in Chapter 1, each character on a keyboard has an ASCII value. Each character pressed is converted into a digital signal, which the computer interprets.

They are a relatively slow method of data entry and are also prone to errors, however keyboards are probably still the easiest way to enter text into a computer. Unfortunately, frequent use of these devices can lead to injuries, such as **repetitive strain injury (RSI)** in the hands and wrists.

Ergonomic keyboards can help to overcome this problem – these have the keys arranged differently as shown in Figure 3.24. They are also designed to give more support to the wrists and hands when doing a lot of typing.

The following diagram (Figure 3.25) and description summarises how the computer recognises a letter pressed on the keyboard:

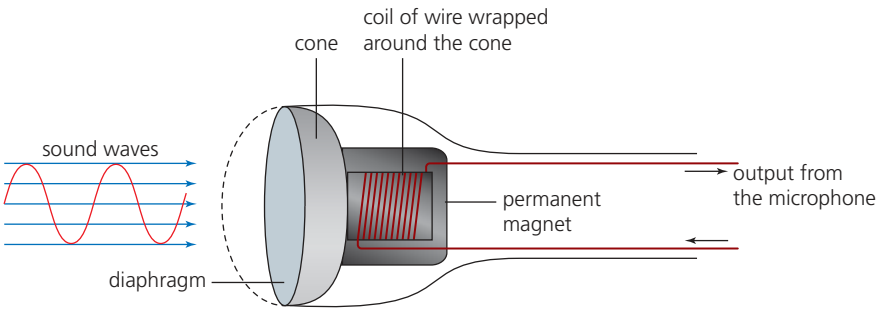
- There is a membrane or circuit board at the base of the keys
- In Figure 3.25, the 'H' key is pressed and this completes a circuit as shown
- The CPU in the computer can then determine which key has been pressed
- The CPU refers to an index file to identify which character the key press represents
- Each character on a keyboard has a corresponding ASCII value (see Chapter 1).



▲ **Figure 3.25** Diagram of a keyboard

Microphones

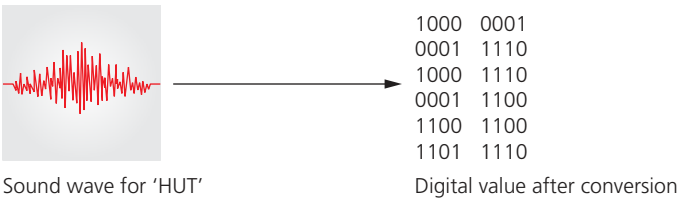
Microphones are either built into the computer or are external devices connected through the USB port or using Bluetooth connectivity. Figure 3.26 shows how a microphone can convert sound waves into an electric current. The current produced is converted to a digital format so that a computer can process it or store it (on, for example, a CD).



▲ **Figure 3.26** Diagram of how a microphone works

- When sound is created, it causes the air to vibrate.
- When a diaphragm in the microphone picks up the air vibrations, the diaphragm also begins to vibrate.
- A copper coil is wrapped around the cone which is connected to the diaphragm. As the diaphragm vibrates, the cone moves in and out causing the copper coil to move backwards and forwards.
- This forwards and backwards motion causes the coil to cut through the magnetic field around the permanent magnet, inducing an electric current.
- The electric current is then either amplified or sent to a recording device. The electric current is analogue in nature.

The electric current output from the microphone can also be sent to a computer where a sound card converts the current into a digital signal which can then be stored in the computer. The following diagram shows what happens when the word 'hut' is picked up by a microphone and is converted into digital values:



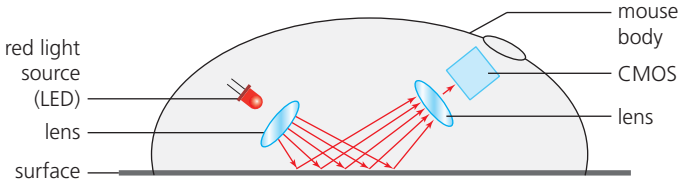
▲ **Figure 3.27** Analogue to digital conversion

Look at Figure 3.27. The word 'hut' (in the form of a sound wave) has been picked up by the microphone; this is then converted using an analogue to digital converter (ADC) into digital values which can then be stored in a computer or manipulated as required using appropriate software.

Optical mouse

An **optical mouse** is an example of a **pointing device**. It uses tiny cameras to take 1500 images per second. Unlike an older mechanical mouse, the optical mouse can work on virtually any surface.

A red LED is used in the base of the mouse and the red light is bounced off the surface and the reflection is picked up by a **complementary metal oxide semiconductor (CMOS)**. The CMOS generates electric pulses to represent the reflected red light and these pulses are sent to a **digital signal processor (DSP)**. The processor can now work out the coordinates of the mouse based on the changing image patterns as it is moved about on the surface. The computer can then move the on-screen cursor to the coordinates sent by the mouse.



▲ **Figure 3.28** Diagram of an optical mouse

Benefits of an optical mouse over a mechanical mouse

- There are no moving parts, therefore it is more reliable.
- Dirt can't get trapped in any of the mechanical components.
- There is no need to have any special surfaces.

Most optical mice use Bluetooth connectivity rather than using a USB wired connection. While this makes the mouse more versatile, a wired mouse has the following advantages:

- no signal loss since there is a constant signal pathway (wire)
- cheaper to operate (no need to buy new batteries or charge batteries)
- fewer environmental issues (no need to dispose of old batteries).

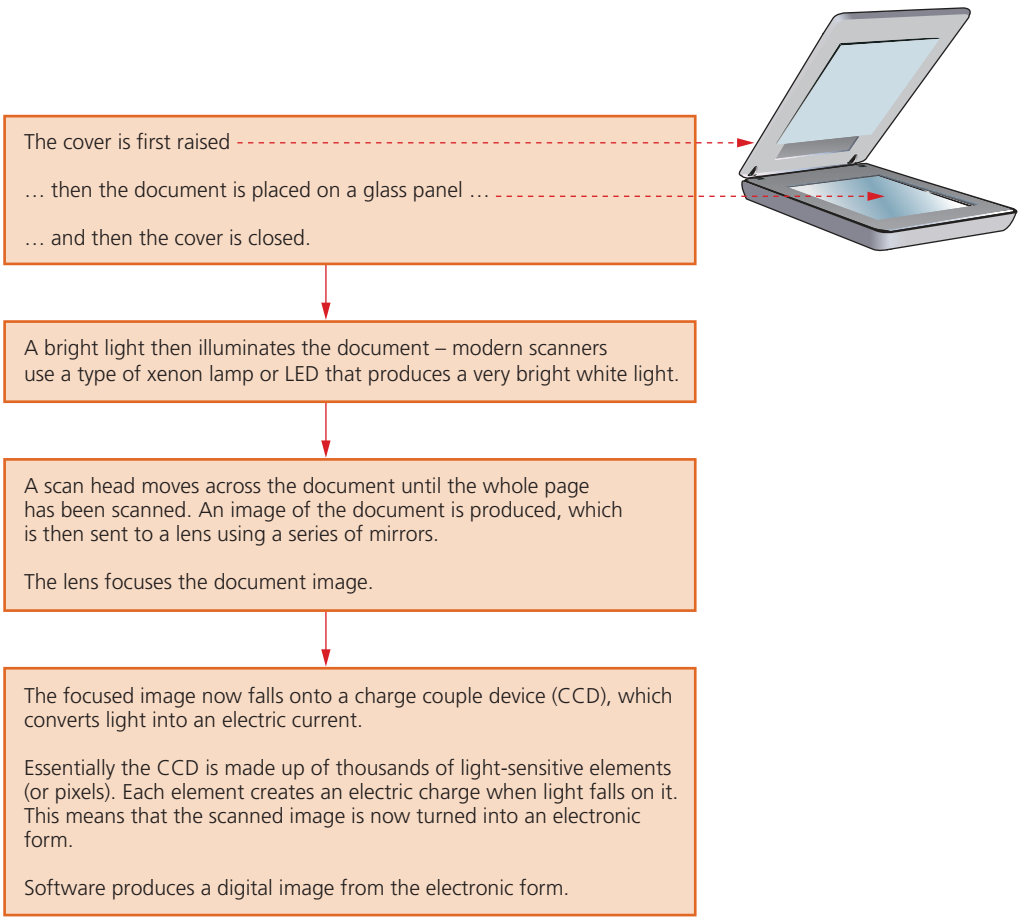
2D and 3D scanners

Scanners are either two dimensional (2D) or three dimensional (3D).

2D scanners

These types of scanner are the most common form and are generally used to input hard copy (paper) documents. The image is converted into an electronic form that can be stored in a computer.

A number of stages occur when scanning a document:



▲ **Figure 3.29** How a 2D scanner works

Computers equipped with **optical character recognition (OCR)** software allow the scanned text from the document to be converted into a **text file format**. This means the scanned image can now be edited and manipulated by importing it into a word processor.

If the original document was a photograph or image, then the scanned image forms an image file such as JPEG.

3D scanners

3D scanners scan solid objects and produce a three-dimensional image. Since solid objects have x, y and z coordinates, these scanners take images at several points along these three coordinates. A digital image which represents the solid object is formed.

The scanned images can be used in **computer aided design (CAD)** or, more recently, sent to a 3D printer (see Section 3.2.2) to produce a working model of the scanned image.

There are numerous technologies used in 3D scanners – lasers, magnetic resonance, white light, and so on. It is beyond the scope of this book to look at these in any great depth; however, the second application that follows describes the technology behind one form of 3D scanning.

Application of 2D scanners at an airport

2D scanners are used at airports to read passports. They make use of OCR technology to produce digital images which represent the passport pages. Because of the OCR technology, these digital images can be manipulated in a number of ways.

Link

For more on ASCII, please see Chapter 1.

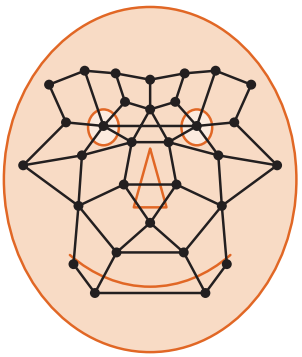
For example, the OCR software is able to review these images, select the text part, and then automatically put the text into the correct fields of an existing database. It is possible for the text to be stored in an ASCII format – it all depends on how the data is to be used.

At many airports the two-dimensional photograph in the passport is scanned and stored as a JPEG image. The passenger's face is also photographed using a digital camera (a 2D image is taken so it can be matched to the image taken from the passport). The two digital images are compared using face recognition/detection software. Key parts of the face are compared.

The face in Figure 3.30 shows several of the positions used by the face recognition software. Each position is checked when the software tries to compare two facial images. Data, such as:

- distance between the eyes
- width of the nose
- shape of the cheek bones
- length of the jaw line
- shape of the eyebrows,

are all used to uniquely identify a given face.



▲ **Figure 3.30** Face recognition

When the image from the passport and the image taken by the camera are compared, these key positions on the face determine whether or not the two images represent the same face.

Application of 3D scanning – computed tomographic (CT) scanners

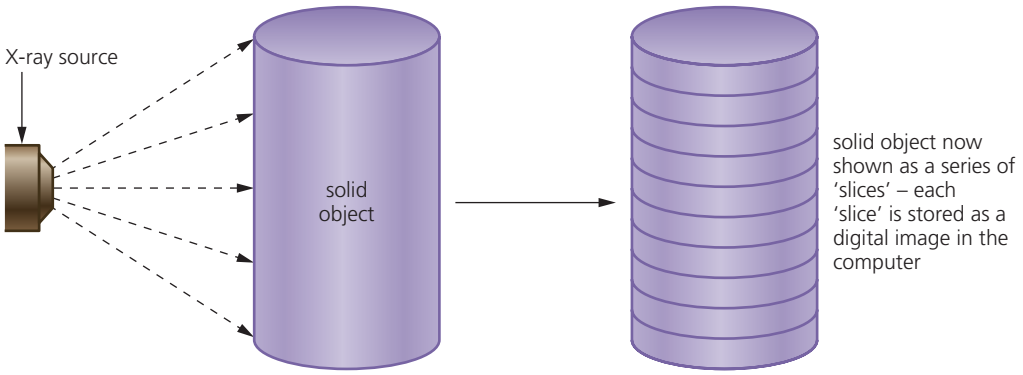
Computed tomographic (CT) scanners are used to create a 3D image of a solid object. This is based on tomography technology, which basically builds up an image of the solid object through a series of very thin 'slices'. Each of these 2D 'slices' make up a representation of the 3D solid object.

Each slice is built up by use of X-rays, radio frequencies or gamma imaging; although a number of other methods exist. Each 'slice' is then stored as a digital image in the computer memory. The whole of the solid object is represented digitally in the computer memory.

Depending on how the image is formed, this type of tomographic scanner can have different names. For example:

Name	CT Scanner	MRI	SPECT
Stands for	computerised tomography	magnetic resonance images	single photon emission computer tomography
Uses	X-rays	radio frequencies	gamma rays

Here is a simple example of how tomography works:



▲ **Figure 3.31** Tomography

Touch screens

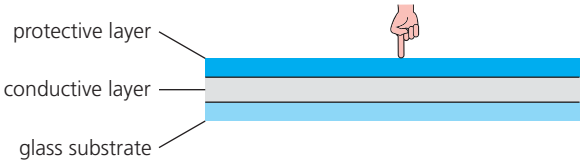
Touch screens are now a very common form of input device. They allow simple touch selection from a menu to launch an application (app). Touch screens allow the user to carry out the same functions as they would with a pointing device, such as a mouse. There are three common types of touch screen technologies currently being used by mobile phone and tablet manufacturers. Similar technologies are used in other touch screen applications (for example, food selection at a fast food restaurant):

- capacitive
- infrared
- resistive (most common method at the moment).

Capacitive touch screens

Capacitive touch screens are composed of a layer of glass (protective layer), a transparent electrode (conductive) layer and a glass substrate (see Figure 3.32). Since human skin is a conductor of electricity, when bare fingers (or a special stylus) touch the screen, the electrostatic field of the conductive layer is

changed. The installed microcontroller is able to calculate where this change took place and hence determine the coordinates of the point of touching.



▲ **Figure 3.32** Capacitive touch screen

There are presently two main types of capacitive touch screens:

- surface
- projective.

The two methods work in a slightly different way but they both have the same general structure as shown in Figure 3.32.

With **surface capacitive screens**, sensors are placed at the corners of a screen. Small voltages are also applied at the corners of the screen creating an electric field. A finger touching the screen surface will draw current from each corner reducing the capacitance. A microcontroller measures the decrease in capacitance and hence determines the point where the finger touched the screen. This system only works with a bare finger or stylus.

Projective capacitive screens work slightly differently to surface capacitive screens. The transparent conductive layer is now in the form of an X-Y matrix pattern. This creates a three dimensional (3D) electrostatic field. When a finger touches the screen, it disturbs the 3D electrostatic field allowing a microcontroller to determine the coordinates of the point of contact. This system works with bare fingers, stylus and thin surgical or cotton gloves. It also allows multi-touch facility (for example, pinching or sliding).

Advantages compared to the other two technologies

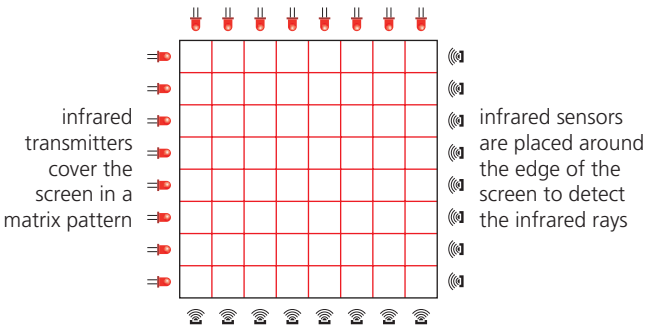
- Better image clarity than resistive screens, especially in strong sunlight
- Very durable screens that have high scratch resistance
- Projective capacitive screens allow multi-touch.

Disadvantages compared to the other two technologies

- Surface capacitive screens only work with bare fingers or a special stylus
- They are sensitive to electromagnetic radiation (such as magnetic fields or microwaves).

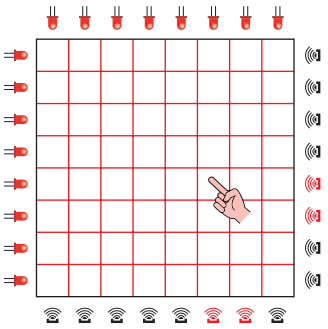
Infrared touch screens

Infrared touch screens use a glass screen with an array of sensors and infrared transmitters.



▲ **Figure 3.33** Array of infrared transmitters and sensors surrounding the screen

The sensors detect the infrared radiation. If any of the infrared beams are broken (for example, with a finger touching the screen), the infrared radiation reaching the sensors is reduced. The sensor readings are sent to a microcontroller that calculates where the screen was touched:



◀ **Figure 3.34** Infrared screen touched causing sensors (shown in red) to show a reduction in infrared radiation – thus the exact position where the screen was touched can be calculated

Advantages compared to the other two technologies

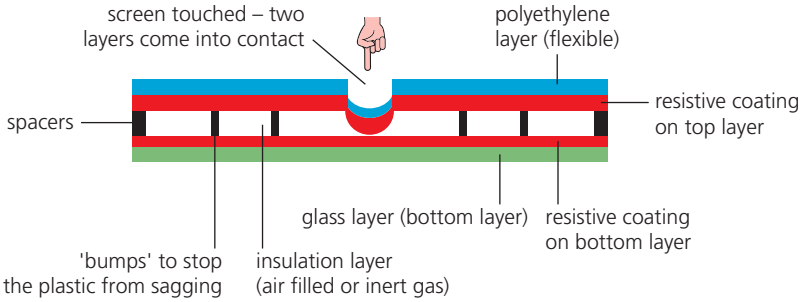
- Allows multi-touch facilities
- Has good screen durability
- The operability isn't affected by a scratched or cracked screen.

Disadvantages compared to the other two technologies

- The screen can be sensitive to water or moisture
- It is possible for accidental activation to take place if the infrared beams are disturbed in some way
- Sometimes sensitive to light interference.

Resistive touch screens

Resistive touch screens are made up of two layers of electrically resistive material with a voltage applied across them. The upper layer is made of flexible polyethylene (a type of polymer) with a resistive coating on one side (see Figure 3.35). The bottom layer is made of glass also with a resistive coating (usually indium tin oxide) on one side. These two layers are separated by air or an inert gas (such as argon). When the top polyethylene surface is touched, the two layers make contact. Since both layers are coated in a resistive material a circuit is now completed which results in a flow of electricity. The point of contact is detected where there was a change in voltage.



▲ **Figure 3.35** Resistive touch screen

A microcontroller converts the voltage (created when the two resistive layers touch) to digital data, which it then sends to the microprocessor.

Advantages compared to the other two technologies

- Good resistance to dust and water
- Can be used with bare fingers, stylus and gloved hand.

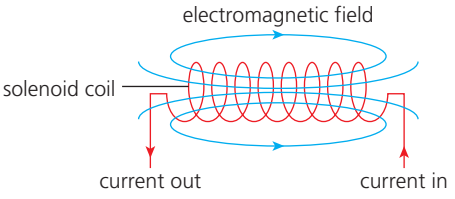
Disadvantages compared to the other two technologies

- Low touch sensitivity (sometimes have to press down harder)
- Doesn't support multi-touch facility
- Poor visibility in strong sunlight
- Vulnerable to scratches on the screen (made of polymer).

3.2.2 Output devices

Actuators

When a computer is used to control devices, such as a conveyer belt or a valve, it is usually necessary to use an **actuator** to, for example, start/stop the conveyer belt or open/close the valve. An actuator is a mechanical or electromechanical device such as a relay, solenoid or motor. We will consider a solenoid as the example; this converts an electrical signal into a magnetic field producing linear motion:



▲ **Figure 3.36** A solenoid

If a plunger (for example, a magnetised metal bar) is placed inside the coil, it will move when a current is applied to the coil (see Figure 3.36). This would allow the solenoid to operate a valve or a switch, for example. There are also examples of rotary solenoids where a cylindrical coil is used. In this case, when a current is supplied to the coil, it would cause a rotational movement of the plunger.

Light projectors

There are two common types of light projector:

- digital light projector (DLP)
- liquid crystal display (LCD) projector.

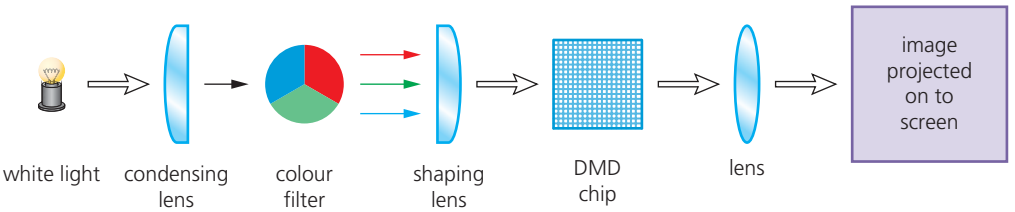
Projectors are used to project computer output onto larger screens or even onto interactive whiteboards. They are often used in presentations and in multimedia applications. The next section compares the basic operation of the two projector technologies.

Digital light projectors (DLP)

The use of millions of micro mirrors on a small **digital micromirror device** (**DMD** chip) is the key to how these devices work.

The number of micro mirrors and the way they are arranged on the DMD chip determines the resolution of the digital image. When the micro mirrors tilt towards the light source, they are ON. When the micro mirrors tilt away from the light source, they are OFF. This creates a light or dark pixel on the projection screen. The micro mirrors can switch on or off several thousand times a second creating various grey shades – typically 1024 grey shades can be produced (for example, if the mirror switches on more often than it switches off, it will produce a lighter shade of grey). This is known as a greyscale image.

A bright white light source (for example, from a xenon bulb) passes through a colour filter on its way to the DMD chip. The white light is split into the primary colours: red, green and blue – the DLP projector can create over 16 million different colours. The ON and OFF states of each micro mirror are linked with colours from the filter to produce the coloured image.



▲ **Figure 3.37** A digital light projector (DLP)

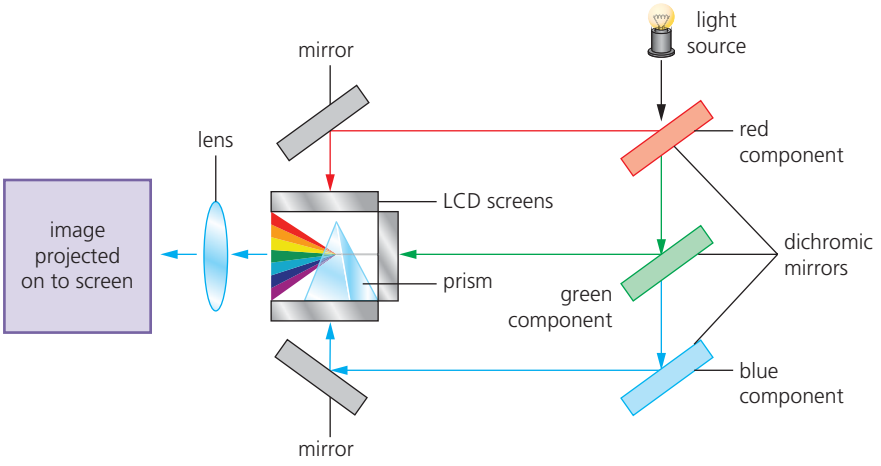
Note: The DMD chip is a microoptoelectromechanical system (MOEMS) that contains several thousand microscopic mirrors (made out of polished aluminium metal) arranged on the chip surface. They are each about 16 µm (16×10^{-6} metres) in size and each corresponds to a pixel in the displayed screen image.

Liquid crystal display (LCD) projector

These are older technology than DLP. Essentially a high-intensity beam of light passes through an LCD display and then onto a screen. How this works in principle is described below:

- a powerful beam of white light is generated from a bulb or LED inside the projector body
- this beam of light is then sent to a group of chromatic-coated mirrors (known as dichromic mirrors); these reflect the light back at different wavelengths

- when the white light hits these mirrors, the reflected light has wavelengths corresponding to red, green and blue light components
- these three different coloured light components pass through three LCD screens (each screen is composed of thousands of tiny pixels which can either block light or let it through; this produces a ***monochromatic*** image)...
- ... consequently, three different versions of the same image are now produced – one is the whole image in different shades of red, one is the whole image in different shades of green and one is the whole image in different shades of blue
- these images are then re-combined using a special prism to produce a full colour image
- finally, the image passes through the projector lens onto a screen.



▲ **Figure 3.38** LCD projector

Advantages and disadvantages of the two types of projector

▼ **Table 3.5** Advantages and disadvantages of DLP and LCD projectors

	Advantages	Disadvantages
Digital light projector (DLP)	higher contrast ratios	image tends to suffer from 'shadows' when showing a moving image
	higher reliability/longevity	
	quieter running than LCD projector	DLP do not have grey components in the image
	uses a single DMD chip, which mean no issues lining up the images	the colour definition is frequently not as good as LCD projectors because the colour saturation is not as good (colour saturation is the intensity of a colour)
	smaller and lighter than LCD projector	
	they are better suited to dusty or smoky atmospheres than LCD projectors	
LCD projector	give a sharper image than DLP projectors	although improving, the contrast ratios are not as good as DLPs
	have better colour saturation than DLP projectors	LCD projectors have a limited life (that is, the longevity is not as good as DLPs)
	more efficient in their use of energy than DLP technology – consequently they generate less heat	since LCD panels are organic in nature, they tend to degrade with time (screens turn yellow and the colours are subsequently degraded over time)

▲ **Figure 3.39** Inkjet printer

Inkjet and laser printers

Inkjet printers

Inkjet printers are essentially made up of:

- a print head, which consists of nozzles that spray droplets of ink onto the paper to form characters
- an ink cartridge or cartridges; either one cartridge for each colour (blue, yellow and magenta) and a black cartridge or one single cartridge containing all three colours + black (Note: some systems use six colours)
- a stepper motor and belt, which moves the print head assembly across the page from side to side
- a paper feed, which automatically feeds the printer with pages as they are required.

The ink droplets are produced currently using two different technologies:

Thermal bubble – tiny resistors create localised heat which makes the ink vaporise. This causes the ink to form a tiny bubble; as the bubble expands, some of the ink is ejected from the print head onto the paper. When the bubble collapses, a small vacuum is created which allows fresh ink to be drawn into the print head. This continues until the printing cycle is completed.

Piezoelectric – a crystal is located at the back of the ink reservoir for each nozzle. The crystal is given a tiny electric charge which makes it vibrate. This vibration forces ink to be ejected onto the paper; at the same time more ink is drawn in for further printing.

When a user wishes to print a document using an inkjet printer, the following sequence of events takes place. Whatever technology is used, the basic steps in the printing process are the same.

▼ **Table 3.6** Steps in inkjet printing process

Stage in process	Description of what happens
1	the data from the document is sent to a printer driver
2	the printer driver ensures that the data is in a format that the chosen printer can understand
3	a check is made by the printer driver to ensure that the chosen printer is available to print (e.g. is it busy, is it off-line, is it out of ink, and so on)
4	the data is then sent to the printer and it is stored in a temporary memory known as a printer buffer
5	a sheet of paper is then fed into the main body of the printer; a sensor detects whether paper is available in the paper feed tray – if it is out of paper (or the paper is jammed) then an error message is sent back to the computer
6	as the sheet of paper is fed through the printer, the print head moves from side to side across the paper printing the text or image; the four ink colours are sprayed in their exact amounts to produce the desired final colour
7	at the end of each full pass of the print head, the paper is advanced very slightly to allow the next line to be printed; this continues until the whole page has been printed
8	if there is more data in the printer buffer, then the whole process from stage 5 is repeated until the buffer is finally empty
9	once the printer buffer is empty, the printer sends an interrupt to the CPU in the computer; this is a request for more data to be sent to the printer; the whole process continues until the whole of the document has been printed

▲ **Figure 3.40** Laser printer

Laser printers

Laser printers use dry powder ink rather than liquid ink and make use of the properties of static electricity to produce the text and images. Unlike inkjet printers, laser printers print the whole page in one go. Colour laser printers use 4 toner cartridges – blue, cyan, magenta and black. Although the actual technology is different to monochrome printers, the printing method is similar but coloured dots are used to build up the text and images.

The following table describes briefly the stages that occur when a document is printed using a laser printer:

▼ **Table 3.7** Steps in laser printing process

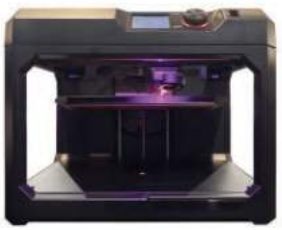
Stage in process	Description of what happens
1	the data from the document is sent to a printer driver
2	the printer driver ensures that the data is in a format that the chosen printer can understand
3	a check is made by the printer driver to ensure that the chosen printer is available to print (e.g. is it busy, is it off-line, is it out of ink, and so on)
4	the data is then sent to the printer and it is stored in a temporary memory known as a printer buffer
5	the start of the printing process involves a printing drum being given a positive charge; as this drum rotates, a laser beam is scanned across it removing the positive charge in certain areas; this leaves negatively charged areas that exactly match the text/images of the page to be printed
6	the drum is then coated with positively charged toner (powdered ink); since the toner is positively charged, it only sticks to the negatively charged parts of the drum
7	a negatively charged sheet of paper is then rolled over the drum
8	the toner on the drum now sticks to the paper to produce an exact copy of the page sent to the printer
9	to prevent the paper sticking to the drum, the electric charge on the paper is removed after one rotation of the drum
10	the paper finally goes through a fuser which is a set of heated rollers; the heat melts the ink so that it fixes permanently to the paper
11	at the very end, a discharge lamp removes all the electric charge from the drum making it ready to print the next page

Applications of inkjet and laser printers

The choice of whether to use an inkjet printer or a laser printer depends on which features make it the most appropriate output device for the given application.

Inkjet printer – inkjet printers are often used for printing one-off photos or where only a few pages of good quality, colour printing is needed; the small ink cartridges or small paper trays would not be an issue with such applications.

Laser printer – these devices produce high quality printouts and are very fast when making multiple copies of a document; any application that needs high volume printing (in colour or monochrome) would choose the laser printer (for example, producing a large number of high-quality flyers or posters for advertising). Laser printers have two advantages: they have large toner cartridges and large paper trays (often holding more than a ream of paper).



▲ **Figure 3.41** Typical 3D printer

▲ **Figure 3.42** An alloy wheel

3D printers

3D printers are used to produce solid objects that actually work. They are primarily based on inkjet and laser printer technology. The solid object is built up layer by layer using materials such as: powdered resin, powdered metal, paper or ceramic.

The alloy wheel in Figure 3.42 was made using an industrial 3D printer.

It was made from many layers (0.1 mm thick) of powdered metal using a technology known as **binder 3D printing**.

Other examples are discussed below.

The following information describes some of the features of 3D printing:

- Various types of 3D printers exist; they range from the size of a microwave oven up to the size of a small car.
- 3D printers use additive manufacturing (i.e. the object is built up layer by layer); this is in sharp contrast to the more traditional method of subtractive manufacturing (i.e. removal of material to make the object). For example, making a statue using a 3D printer would involve building it up layer by layer using powdered stone until the final object was formed. The subtractive method would involve carving the statue out of solid stone (i.e. removing the stone not required) until the final item was produced. Similarly, CNC machining removes metal to form an object; 3D printing would produce the same item by building up the object from layers of powdered metal.
- **Direct 3D printing** uses inkjet technology; a print head can move left to right as in a normal printer. However, the print head can also move up and down to build up the layers of an object.
- **Binder 3D printing** is similar to direct 3D printing. However, this method uses two passes for each of the layers; the first pass sprays dry powder and then on the second pass a binder (a type of glue) is sprayed to form a solid layer.
- Newer technologies are using lasers and UV light to harden liquid polymers; this further increases the diversity of products which can be made.

How to create a solid object using 3D printers

There are a number of steps in the process of producing an object using 3D printers. The steps are summarised below:

A design is made using computer aided design (CAD) software

↓

The finalised drawing is imported into some special software that prepares it in a format that is understood by the 3D printer

↓

The 3D printer is first set up to allow the solid object to be made

↓

The solid object is built up layer by layer (often only 0.1 mm thick); this can take several hours depending on the thickness of the layers, the material used and the size of the final object

↓

The object is removed from the printer and is then prepared; for example, some use a jelly-like support that needs to be washed away by immersion in water, some require the removal of excess plastic powder and others require the cutting away of unused material; in many cases, the object has to be left to 'cure' for a few hours.

◀ **Figure 3.43** How to create an object using a 3D printer

Uses of 3D printing

3D printing is regarded as being possibly the next 'industrial revolution' since it will change the manufacturing methods in many industries. The following list is just a glimpse into what we know can be made using these printers; in the years that follow, this list will probably fill an entire book:

- the covering of prosthetic limbs can be made to exactly fit the limb
- making items to allow precision reconstructive surgery (e.g. facial reconstruction following an accident); the parts made by this technique are more precise in their design since they can be made from exact scanning of the skull
- in aerospace, manufacturers are looking at making wings and other parts using 3D technology; the bonus will be lightweight precision parts
- fashion and art – 3D printing allows new creative ideas to be developed
- making parts for items no longer in production e.g. suspension parts for a vintage car.

These are just a few of the exciting applications which make use of this new technology.

Find out more

The reader is invited to do a search on the internet to find out new and innovative research into 3D printing applications.

LED and LCD screens

LED screens

An LED screen is made up of tiny light emitting diodes (LEDs). Each LED is either red, green or blue in colour. By varying the electric current sent to each LED, its brightness can be controlled, producing a vast range of colours.

This type of screen tends to be used for large outdoor displays due to the brilliance of the colours produced. Recent advancements in LED technology have led to the introduction of OLED (organic LED) screens (see later).

The reader needs to be very careful here. Many television screens are advertised as LED when in fact they are LCD screens which are *backlit* using LEDs.

LCD screens

LCD screens are made up of tiny liquid crystals. These tiny crystals make up an array of pixels that are affected by changes in applied electric fields. How this works is outside the scope of this book. But the important thing to realise is that for LCD screens to work, they require some form of backlighting.

Because LCD's don't produce any light, LCD screens are back-lit using light emitting diode (LED) technology and must not be confused with pure LED screens. Use of LED backlighting gives a very good contrast and brightness range. Before the use of LEDs, LCD screens used cold cathode fluorescent lamp (CCFL) as the back-lit method.

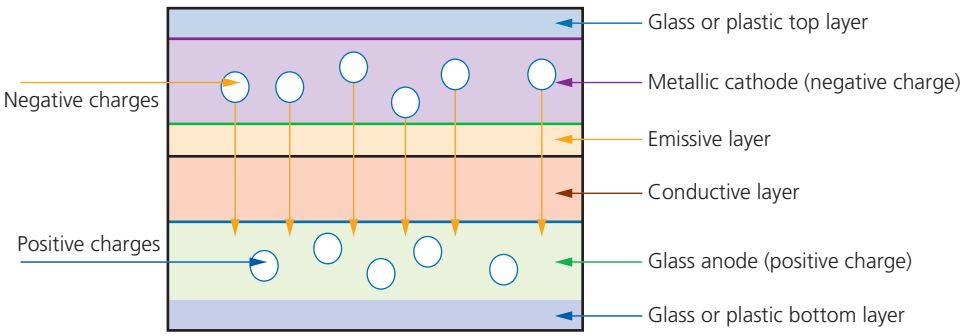
Essentially, CCFL uses two fluorescent tubes behind the LCD screen which supply the light source. When LEDs are used, a matrix of tiny blue-white LEDs is used behind the LCD screen.

LEDs have become increasingly more popular, as the method of back lighting, due to a number of advantages over older CCFL technology:

- LEDs reach their maximum brightness almost immediately (there is no need to 'warm up' before reaching full efficiency)
- LEDs give a whiter light that sharpens the image and makes the colours appear more vivid; CCFL had a slightly yellowish tint
- LEDs produce a brighter light that improves the colour definition
- monitors using LED technology are much thinner than monitors using CCFL technology
- LEDs last indefinitely; this makes the technology more reliable and makes for a more consistent product
- LEDs consume very little power which means they produce less heat as well as using less energy.

Organic light emitting diodes (OLED)

Newer LED technology is making use of **organic light emitting diodes (OLEDs)**. These use organic materials (made up of carbon compounds) to create semi-conductors that are very flexible. Organic films are sandwiched between two charged electrodes (one is a metallic **cathode** and the other a glass **anode**). When an electric field is applied to the electrodes, they give off light. This means that no form of backlighting is required. This allows for very thin screens. It also means that there is no longer a need to use LCD technology, since OLED is a self-contained system.



▲ **Figure 3.44** How an OLED screen works

▲ **Figure 3.45** OLED television (curved screen)

But the important aspect of OLED technology is how thin this makes the screen. It is possible, using OLED technology, to bend screens to any shape (see Figure 3.45). When this is adopted by mobile phone manufacturers, it makes it possible to develop phones that can wrap around your wrist – much like a watch strap. Imagine screens so thin that they can be folded up and placed in your pocket until they are needed. Or how about using folding OLED displays attached to fabrics creating 'smart' clothing (this could be used on outdoor survival clothing where an integrated circuit, mobile phone, GPS receiver and OLED display could all be sewn into the clothing)?

Advantages of using OLED compared to existing LEDs and LCDs:

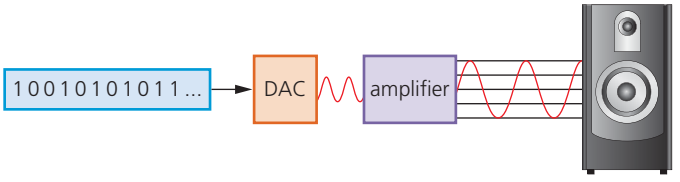
- The plastic, organic layers of an OLED are thinner, lighter and more flexible than the crystal structures used in LEDs or LCDs.
- The light-emitting layers of an OLED are lighter; OLED layers can be made from plastic rather than the glass as used in LED and LCD screens.
- OLEDs give a brighter light than LEDs.
- OLEDs do not require backlighting like LCD screens – OLEDs generate their own light.
- Since OLEDs require no backlighting, they use much less power than LCD screens (most of the LCD power is used to do the backlighting); this is very important in battery-operated devices such as mobile phones.
- Since OLEDs are essentially plastics, they can be made into large, thin sheets (this means they could be used on large advertising boards in airports, subways, and so on).
- OLEDs have a very large field of view, about 170 degrees, which makes them ideal for use in television sets and for advertising screens.

(Loud) speakers

Loudspeakers are output devices that produce sound. When connected to a computer system, digitised sound stored on a file needs to be converted into sound as follows:

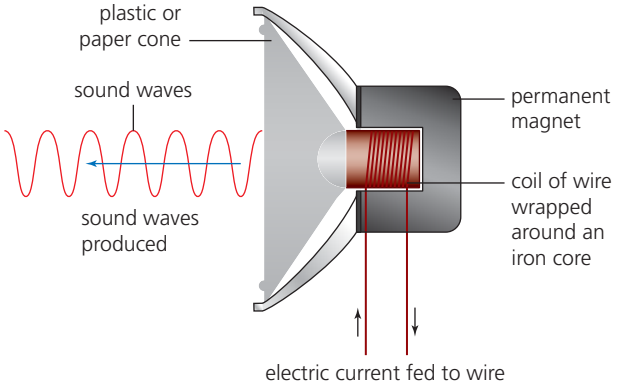
- The digital data is first passed through a digital to analogue converter (DAC) where it is changed into an electric current.
- This is then passed through an amplifier (since the current generated by the DAC will be very small); this creates a current large enough to drive a loudspeaker.
- This electric current is then fed to a loudspeaker where it is converted into sound.

The following schematic shows how this is done:



▲ **Figure 3.46** Digital to analogue conversion

As Figure 3.46 shows, if the sound is stored in a computer file, it must pass through a **digital to analogue converter (DAC)** to convert binary (digital) data into an analogue form (electric current) that can then drive the loudspeaker. Figure 3.47 shows how the loudspeaker converts the electric current into sound:



▲ **Figure 3.47** Diagram showing how a loudspeaker works

- When an electric current flows through the coil of wire that is wrapped around an iron core, the core becomes a temporary electromagnet; a permanent magnet is also positioned very close to this electromagnet.
- As the electric current through the coil of wire varies, the induced magnetic field in the iron core also varies. This causes the iron core to be attracted towards the permanent magnet and as the current varies this will cause the iron core to vibrate.
- Since the iron core is attached to a cone (made of paper or thin synthetic material), this causes the cone to vibrate, producing sound waves.

Activity 3.4

1 **a** Explain the main differences in operation of a laser printer compared to an inkjet printer.

b **i** Name one application of a laser printer and one application of an inkjet printer.

ii For each of your named applications in **b i**, give a reason why the chosen printer is the most suitable.

2 The nine stages in printing a page using an inkjet printer are shown below. The nine stages are NOT in the correct order.

By writing the letters ***A*** to ***I***, put each of the stages into the correct order.

A – the data is then sent to the printer and it is stored in a temporary memory known as a printer buffer

B – as the sheet of paper is fed through the printer, the print head moves from side to side across the paper printing the text or image; the four ink colours are sprayed in their exact amounts to produce the desired final colour

C – the data from the document is sent to a printer driver

D – once the printer buffer is empty, the printer sends an interrupt to the CPU in the computer; this is a request for more data to be sent to the printer; the whole process continues until the whole of the document has been printed

E – the printer driver ensures that the data is in a format that the chosen printer can understand

F – at the end of each full pass of the print head, the paper is advanced very slightly to allow the next line to be printed; this continues until the whole page has been printed

G – a check is made by the printer driver to ensure that the chosen printer is available to print (e.g. is it busy, is it off-line, is it out of ink, and so on)

H – if there is more data in the printer buffer, then the whole process from stage 5 is repeated until the buffer is finally empty

I – a sheet of paper is then fed into the main body of the printer; a sensor detects whether paper is available in the paper feed tray – if it is out of paper (or the paper is jammed) then an error message is sent back to the computer

3 **a** Explain the difference between LED screens and LCD-LED backlit screens.

b Modern LCD screens use blue-white LEDs as backlighting. Cold cathode ray (fluorescent) tubes were used. Give three advantages of using LEDs.

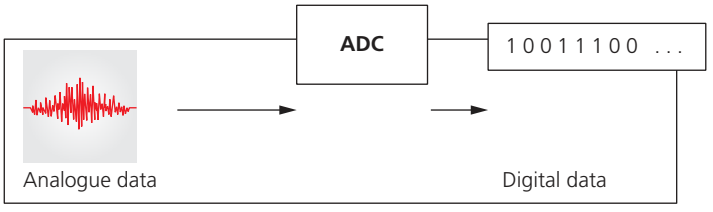
4 Filipe has music stored on his computer's backing store. He wishes to listen to his music through a pair of loudspeakers. Describe how the music, which is digitally stored can be played through his two analogue loudspeakers.

▲ **Figure 3.48** Mercury thermometer

3.2.3 Sensors

Sensors are input devices which read or measure physical properties from their surroundings. Examples include temperature, pressure, acidity level and length (there are many others). Real data is analogue in nature; this means it is constantly changing and doesn't have a single discrete value. Therefore, analogue data needs some form of interpretation by the user, for example, the temperature measurement on a mercury thermometer requires the user to look at the height of the mercury column and use their best judgement (by looking at the scale) to find the temperature. There are an infinite number of values depending on how precisely the height of the mercury column is measured.

However, computers cannot make any sense of these physical quantities so the data needs to be converted into a digital format. This is usually achieved by an **analogue to digital converter (ADC)**. This device converts physical values into discrete digital values.



▲ **Figure 3.49** ADC

Link

For more on actuators see Section 3.2.2.

When the computer is used to control devices, such as a motor or a valve, it is necessary to use a digital to analogue converter (DAC) since these devices need analogue data to operate in many cases. Actuators are used in such control applications.

Sensor readings may cause the microprocessor to, for example, alter a valve or a motor that will then change the next reading taken by the sensor. So the output from the microprocessor will impact on the next input received as it attempts to bring the system within the desired parameters. This is known as feedback.

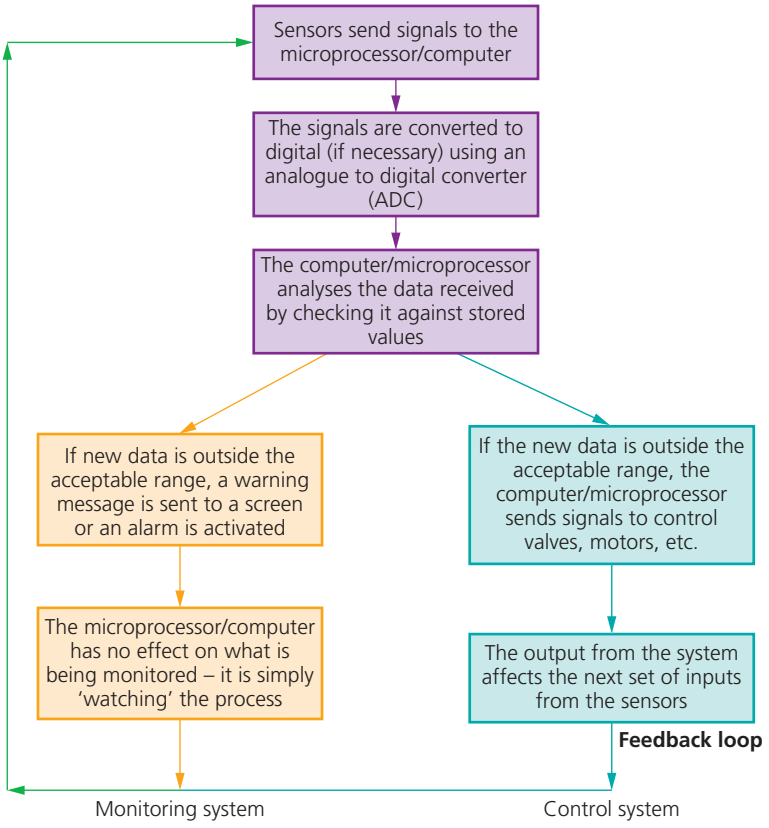
It is important to realise that sensors send out constant values; they don't suddenly send a reading when the parameter they are measuring changes. It is the microprocessor they are giving the input to that will analyse the incoming data and take the necessary action.

Table 3.8 shows a number of common sensors and examples of applications where the sensors might be used.

▼ **Table 3.8** Sensors

Sensor	Description of sensor	Example applications
Temperature	measures temperature of the surroundings by sending signals; these signals will change as the temperature changes	• control of a central heating system • control/monitor a chemical process • control/monitor temperature in a greenhouse
Moisture	measures water levels in, for example, soil (it is based on the electrical resistance of the sample being monitored)	• control/monitor moisture levels in soil in a greenhouse • monitor the moisture levels in a food processing factory
Humidity	this is slightly different to moisture; this measures the amount of water vapour in, for example, a sample of air (based on the fact that the conductivity of air will change depending on the amount of water present)	• monitor humidity levels in a building • monitor humidity levels in a factory manufacturing microchips • monitor/control humidity levels in the air in a greenhouse
Light	these use photoelectric cells that produce an output (in the form of an electric current) depending on the brightness of the light	• switching street lights on or off depending on light levels • switch on car headlights automatically when it gets dark
Infrared (active)	these use an invisible beam of infrared radiation picked up by a detector; if the beam is broken, then there will be a change in the amount of infrared radiation reaching the detector (sensor)	• turn on car windscreen wipers automatically when it detects rain on the windscreen • security alarm system (intruder breaks the infra-red beam)
Infrared (passive)	these sensors measure the heat radiation given off by an object, for example, the temperature of an intruder or the temperature in a fridge	• security alarm system (detects body heat) • monitor the temperature inside an industrial freezer or chiller unit
Pressure	a pressure sensor is a transducer and generates different electric currents depending on the pressure applied	• weighing of lorries at a weighing station • measure the gas pressure in a nuclear reactor
Acoustic/sound	these are basically microphones that convert detected sound into electric signals/pulses	• pick up the noise of footsteps in a security system • detect the sound of liquids dripping at a faulty pipe joint
Gas	most common ones are oxygen or carbon dioxide sensors; they use various methods to detect the gas being monitored and produce outputs that vary with the oxygen or carbon dioxide levels present	• monitor pollution levels in the air at an airport • monitor oxygen and carbon dioxide levels in a greenhouse • monitor oxygen levels in a car exhaust
pH	these measure acidity through changes in voltages in, for example, soil	• monitor/control acidity levels in the soil in a greenhouse • control acidity levels in a chemical process
Magnetic field	these sensors measure changes in magnetic fields – the signal output will depend on how the magnetic field changes	• detect magnetic field changes (for example, in mobile phones and CD players) • used in anti-lock braking systems in cars
Accelerometer	these are sensors that measure acceleration and motion of an application, i.e. the change in velocity (a piezoelectric cell is used whose output varies according to the change in velocity)	• used in cars to measure rapid deceleration and apply air bags in a crash • used by mobile phones to change between portrait and landscape mode
Proximity	these sensors detect the presence of a nearby object	• detect when a face is close to a mobile phone screen and switches off screen when held to the ear
Flow (rate)	these sensors measure the flow rate of a moving liquid or gas and produce an output based on the amount of liquid or gas passing over the sensor	• used in respiratory devices and inhalers in hospitals • measure gas flows in pipes (for example, natural gas)
Level	these sensors use ultrasonics (to detect changing liquid levels in, for example, a tank) or capacitance/conductivity (to measure static levels (for example, height of water in a river) – note, level sensors can also be optical or mechanical in nature	• monitor levels in a petrol tank in a car • in a pharmaceutical process where powder levels in tablet production need to be monitored • leak detection in refrigerant (air conditioning)

Sensors are used in both monitoring and control applications. There is a subtle difference between how these two methods work (the flowchart is a simplification of the process):



▲ **Figure 3.50** Monitoring and control systems using sensors

Examples of monitoring

- Monitoring of a patient in a hospital for vital signs such as heart rate, temperature, etc.
- Monitoring of intruders in a burglar alarm system
- Checking the temperature levels in a car engine
- Monitoring pollution levels in a river.

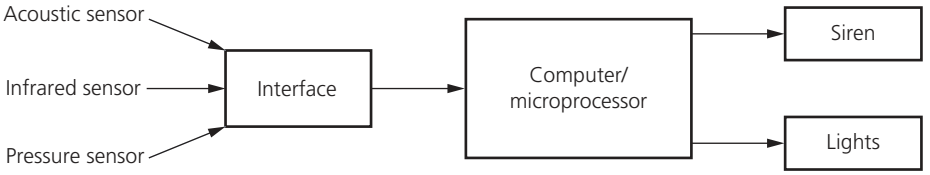
Examples of control

- Turning street lights on at night and turning them off again during daylight
- Controlling the temperature in a central heating/air conditioning system
- Chemical process control (for example, maintaining temperature and pH of process)

- Operating anti-lock brakes on a car when necessary
- Controlling the environment in a green house.

Monitoring applications

Security systems



▲ **Figure 3.51** Security system

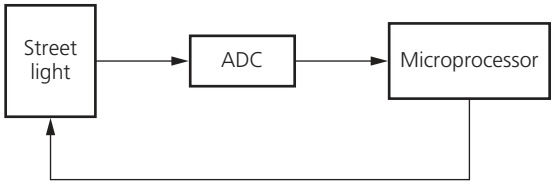
Note: compare this to Figure 3.9 (embedded systems) which shows the security system in more detail. Figure 3.51 concentrates on the sensor input.

The security monitoring system will carry out the following actions:

- the system is activated by keying in a password on a keypad
- the **infrared sensor** will pick up the movement of an intruder in the building
- the **acoustic sensor** will pick up sounds such as footsteps or breaking glass
- the **pressure sensor** will pick up the weight of an intruder coming through a door or through a window
- the sensor data is passed through an ADC if it is in an analogue form ...
- ... to produce digital data
- the computer/microprocessor will sample the digital data coming from these sensors at a given frequency (e.g. every 5 seconds) ...
- ... the data is compared with the stored values by the computer/microprocessor
- if any of the incoming data values are outside the acceptable range, then the computer sends a signal ...
- ... to a siren to sound the alarm, or
- ... to a light to start flashing
- a DAC is used if the devices need analogue values to operate them
- the alarm continues to sound/lights continue to flash until the system is re-set with a password.

Monitoring of patients in a hospital

- A number of sensors are attached to the patient ...
- ... these measure vital signs such as: temperature, heart rate, breathing rate, etc.
- these sensors are all attached to a computer system
- the sensors constantly send data back to the computer system
- the computer samples the data at frequent intervals
- the range of acceptable values for each parameter is keyed into the computer
- the computer compares the values from the sensors with those values keyed in
- if anything is out of the acceptable range, a signal is sent by the computer ...
- ... to sound an alarm
- if data from the sensors is within range, the values are shown in either graphical form on a screen and/or a digital read out
- monitoring continues until the sensors are disconnected from the patient.



▲ **Figure 3.52** Street lighting

Control applications

Control of street lighting

This next sequence shows how a microprocessor is used to control the operation of a street lamp. The lamp is fitted with a light sensor which constantly sends data to the microprocessor. The data value from the sensor changes according to whether it is sunny, cloudy, raining or it is night time (etc.):

- the light sensor sends data to the ADC interface
- this changes the data into digital form and sends it to the microprocessor
- the microprocessor samples the data every minute (or at some other frequency rate)
- if the data from the sensor < value stored in memory …
- … a signal is sent from the microprocessor to the street lamp …
- … and the lamp is switched on
- the lamp stays switched on for 30 minutes before the sensor readings are sampled again (this prevents the lamp flickering off and on during brief heavy cloud cover, for example)
- if the data from the sensor >= value stored in memory …
- … a signal is sent from the microprocessor to the street lamp …
- … and the lamp is switched off
- the lamp stays switched off for 30 minutes before sensor readings are sampled again (this prevents the lamp flickering off and on during heavy cloud cover for example).

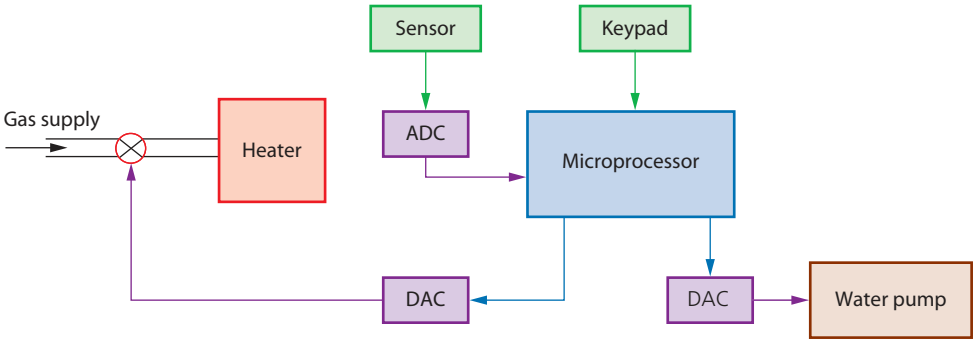
Anti-lock braking systems (on cars)

Anti-lock braking systems (ABS) on cars use **magnetic field sensors** to stop the wheels locking up on the car if the brakes have been applied too sharply:

- when one of the car wheels rotates too slowly (i.e. it is locking up), a magnetic field sensor sends data to a microprocessor
- the microprocessor checks the rotation speed of the other three wheels
- if they are different (i.e. rotating faster), the microprocessor sends a signal to the braking system …
- … and the braking pressure to the affected wheel is reduced …
- … the wheel's rotational speed is then increased to match the other wheels
- the checking of the rotational speed using these magnetic field sensors is done several times a second …
- … and the braking pressure to all the wheels can be constantly changing to prevent any of the wheels locking up under heavy braking …
- … this is felt as a 'judder' on the brake pedal as the braking system is constantly switched off and on to equalise the rotational speed of all four wheels
- if one of the wheels is rotating too quickly, braking pressure is increased to that wheel until it matches the other three.

Central heating systems

In this example, a gas supply is used to heat water using a heater. A valve on the gas supply is controlled by a microprocessor and is opened if the heating levels need to be increased. A water pump is used to pump hot water around the central heating system whenever the temperature drops below a pre-set value:



▲ **Figure 3.53** Controlling a central heating system

So how does this work?

- the required temperature is keyed in and this is stored in the microprocessor memory (this is called the pre-set value)
- the temperature sensor is constantly sending data readings to the microprocessor
- the sensor data is first sent to an ADC to convert the analogue data into digital data
- the digital data is sent to the microprocessor
- the microprocessor compares this data with the pre-set value
- if the temperature reading >= pre-set value then no action is taken
- if the temperature reading < pre-set value, then a signal is sent …
- … to an actuator (via a DAC) to open the gas valve to the heater
- … to an actuator (via a DAC) to turn on the water pump
- the process continues until the central heating is switched off.

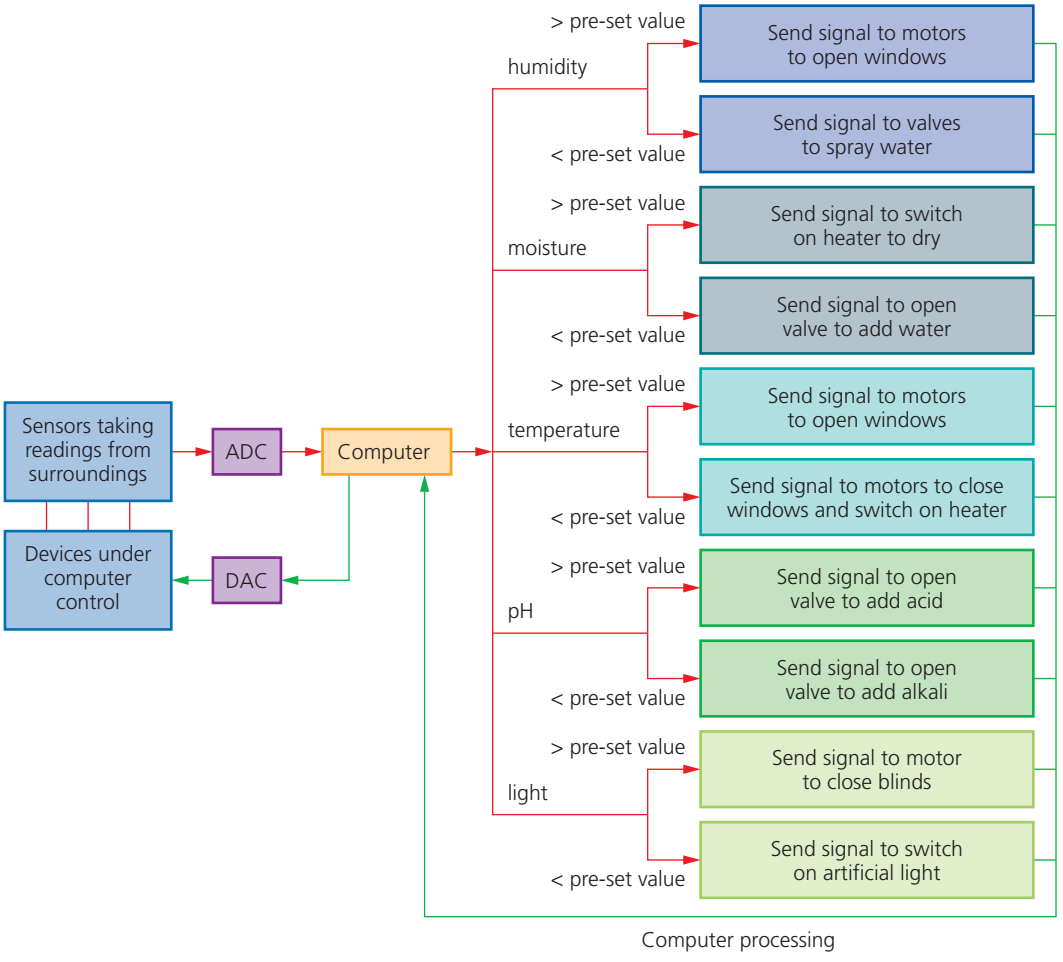
Chemical process control

A certain chemical process only works if the temperature is above 70°C and the pH (acidity) level is less than 3.5. Sensors are used as part of the control system. A heater is used to heat the reactor and valves are used to add acid when necessary to maintain the acidity. The following description shows how the sensors and computer are used to control this process:

- **temperature** and **pH sensors** read data from the chemical process
- this data is converted to digital using an ADC and is then sent to the computer
- the computer compares the incoming data with pre-set values stored in memory
- … if the temperature < 70°C, a signal is sent to switch on the heater
- … if the temperature >= 70°C, a signal is sent to switch off the heaters
- … if the pH > 3.5, then a signal is sent to open a valve and acid is added
- … if the pH <= 3.5, then a signal is sent to close this valve
- the computer signals will be changed into analogue signals using a DAC so that it can control the heaters and valves
- this continues as long as the computer system is activated.

Greenhouse environment control

Five different sensors could be used here to control the greenhouse environment, namely: **humidity**, **moisture**, **temperature**, **pH** and **light**. To simplify this problem the control mechanisms are shown in Figure 3.54.



▲ **Figure 3.54** Control of greenhouse environment

Because of the number of sensors, this is clearly quite a complex problem. Let us consider the humidity sensor only. This sends a signal to an ADC, which then sends a digital signal to the computer. This compares the input with stored (pre-set) values and decides what action needs to be taken (follow the orange lines in Figure 3.54). If humidity is > pre-set value, the computer sends a signal to a DAC (follow the green lines in the figure) to operate the motors to open windows thus reducing the humidity. If it is < pre-set value, the computer sends a signal to open valves to spray water into the air (follow the green lines). If the reading = pre-set value, then no action is taken (this isn't shown in the diagram since it could follow either direction). The control process continues as long as the system is switched on. Similar arguments can be used for all five sensors.

Activity 3.5

1 An air conditioning unit in a car is being controlled by a microprocessor and a number of sensors.

   a Describe the main differences between ***control*** and ***monitoring*** of a process.

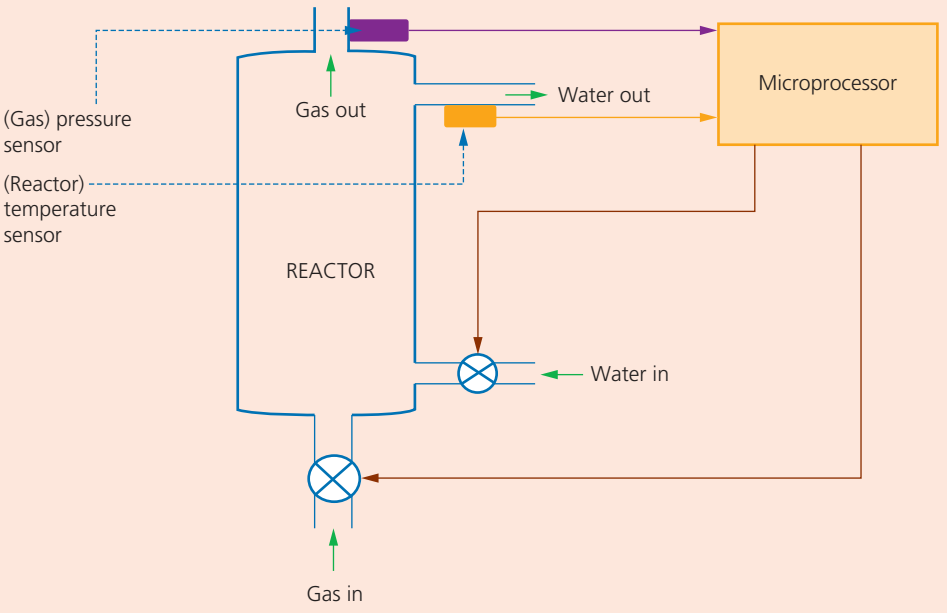
   b Describe how the sensors and microprocessor would be used to control the air conditioning unit in the car. Name at least two different sensors that might be used and explain the role of positive feedback in your description.

   You might find drawing a diagram of your intended process to be helpful.

2 Look at Figure 3.54 and describe how the pH sensor would be used to control the acidity levels in the soil to optimise growing conditions in the greenhouse.

3 The diagram (Figure 3.55) below shows a nuclear reactor. Two of the sensors used in the control and monitoring of the reactor are:

   - a temperature sensor to monitor the reactor temperature (if this exceeds 300°C then the water flow into the reactor is increased)
   - a pressure sensor to monitor the gas pressure of carbon dioxide circulating in the reactor (if this is less 10 bar then the gas pump is opened)
   - note that: ⊗ represents a gas or liquid pump

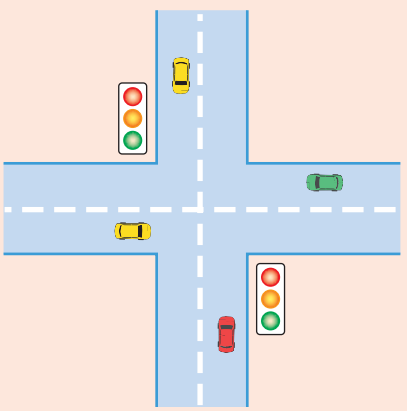


▲ **Figure 3.55**

Describe how the sensors and microprocessor are used to maintain the correct water (reactor) temperature and gas pressure in the reactor. Name any other hardware devices you think may be needed in your description.

4 The junction (Figure 3.56) is controlled by traffic lights.

Describe how sensors in the road and a microprocessor are used to control the traffic at the junction. The microprocessor is able to change the colour sequence of the lights.



▲ **Figure 3.56**

3.3 Data storage

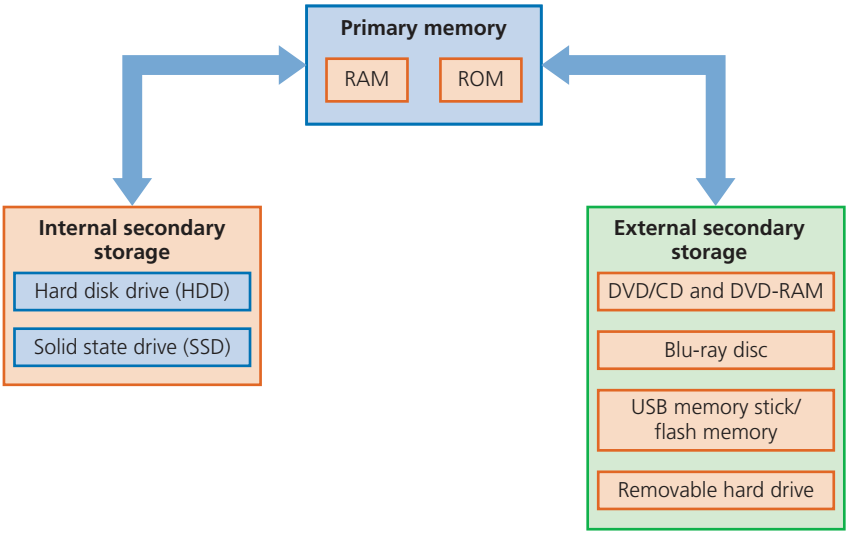
All computers require some form of memory and storage. Memory is usually referred to as the internal devices used to store data that the computer can access directly. This is also known as primary memory. This memory can be the user's workspace, temporary data or data that is key to running the computer.

Storage devices allow users to store applications, data and files. The user's data is stored permanently and they can change it or read it as they wish. Storage needs to be larger than internal memory since the user may wish to store large files (such as music files or videos). Storage devices can also be removable to allow data, for example, to be transferred between computers. Removable devices allow a user to store important data in a different location in case of data loss.

However, all of this removeable storage has become less important with the advent of technology such as 'data drop' (which uses Bluetooth) and cloud storage. Figure 3.57 summarises the types of memory and storage devices covered in this section.

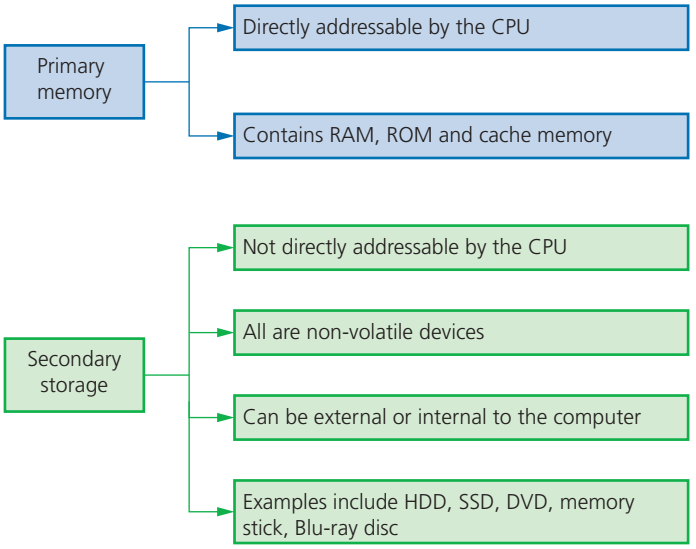
Memory and storage devices can be split up into two distinct groups:

- primary memory
- secondary storage.



▲ **Figure 3.57** Typical memory and storage devices

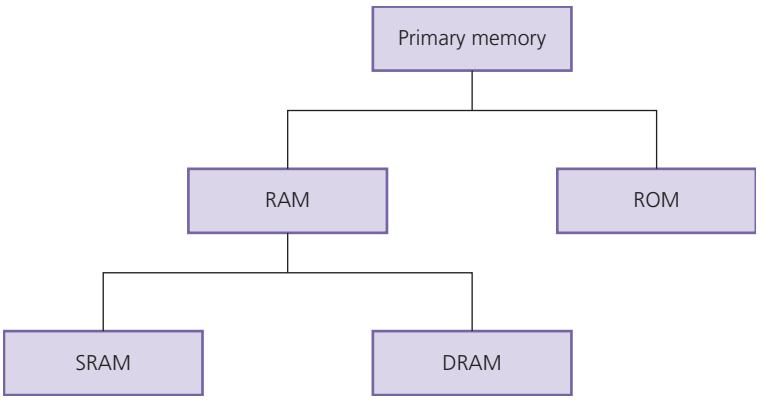
Here is a summary of the differences between primary memory and storage devices:



▲ **Figure 3.58** Summary of primary, secondary and off-line devices

3.3.1 Primary memory

Primary memory is the part of the computer memory which can be accessed directly from the CPU; this includes **random access memory (RAM)** and **read-only memory (ROM)** memory chips. Primary memory allows the CPU to access applications and services temporarily stored in memory locations. The structure of primary memory is shown in Figure 3.59.



▲ **Figure 3.59** Primary memory

Random access memory (RAM)

All computer systems come with some form of RAM. These memory devices are not really random; this refers to the fact that any memory location in RAM can be accessed independent of which memory location was last used. When you run an application or program, data is retrieved from secondary storage and placed temporarily into RAM. Access time to locate data is much faster in RAM than in secondary or off-line devices. Features of RAM include:

- can be written to or read from, and the data can be changed by the user or the computer (i.e. it is a temporary memory)

- used to store data, files, part of an application or part of the operating system **currently in use**
- it is **volatile**, which means memory contents are lost when powering off the computer.

In general, the larger the size of RAM the faster the computer will operate. In reality, RAM never runs out of memory; it continues to operate but just becomes slower and slower as more data is stored. As RAM becomes 'full', the CPU has to continually access the secondary data storage devices to overwrite **old** data on RAM with **new** data. By increasing the RAM size, the number of times this has to be done is considerably reduced; thus making the computer operate more quickly.

There are currently two types of RAM technology:

- dynamic RAM (DRAM)
- static RAM (SRAM).

Dynamic RAM (DRAM)

▲ **Figure 3.60** DRAM

▲ **Figure 3.61** SRAM

Each DRAM chip consists of transistors and capacitors. Each of these parts is tiny since a single RAM chip will contain millions of transistors and capacitors. The function of each part is:

- capacitor – this holds the bits of information (0 or 1)
- transistor – this acts like a switch; it allows the chip control circuitry to read the capacitor or change the capacitor's value.

This type of RAM needs to be constantly **refreshed** (that is, the capacitor needs to be re-charged every 15 microseconds otherwise it would lose its value). If it wasn't refreshed, the capacitor's charge would leak away very quickly leaving every capacitor with the value 0.

DRAMs have a number of advantages over SRAMs:

- they are much less expensive to manufacture than SRAM
- they consume less power than SRAM
- they have a higher memory capacity than SRAM.

Static RAM (SRAM)

A major difference between SRAM and DRAM is that SRAM doesn't need to be constantly refreshed.

It makes use of **flip flops**, which hold each bit of memory.

SRAM is much faster than DRAM when it comes to data access (typically, access time for SRAM is 25 nanoseconds and for DRAM is 60 nanoseconds).

DRAM is the most common type of RAM used in computers, but where absolute speed is essential, for example, in the CPU's memory cache, SRAM is the preferred technology. Memory cache is a high-speed portion of the memory; it is effective because most programs access the same data or instructions many times. By keeping as much of this information as possible in SRAM, the computer avoids having to access the slower DRAM.

Table 3.9 summarises the differences between DRAM and SRAM.

▼ **Table 3.9** Differences between DRAM and SRAM

DRAM	SRAM
consists of a number of transistors and capacitors	uses flip flops to hold each bit of memory
needs to be constantly refreshed	doesn't need to be constantly refreshed
less expensive to manufacture than SRAM	has a faster data access time than DRAM
has a higher memory capacity than SRAM	CPU memory cache makes use of SRAM
main memory is constructed from DRAM	
consumes less power than SRAM	

Read-only memory (ROM)

Another form of primary memory is read-only memory (ROM). This is similar to RAM in that it shares some of its properties, but the main difference is that it cannot be changed or written to. ROM chips have the following features:

- they are non-volatile (the contents are not lost after powering off the computer)
- they are permanent memories (the contents cannot be changed or written to by the user, the computer or any application/program)
- the contents can only be read
- they are often used to store data that the computer needs to access when powering up for the first time (the basic input/output system (BIOS)); these are known as the start-up instructions (or bootstrap)

Here is a summary of the main differences between RAM and ROM:

▼ **Table 3.10** RAM and ROM features

RAM	ROM
temporary memory device	permanent memory device
volatile memory	non-volatile memory device
can be written to and read from	data stored cannot be altered
used to store data, files, programs, part of OS **currently** in use	always used to store BIOS and other data needed at start up
can be increased in size to improve operational speed of a computer	

Example of an application

We will now consider an application, other than a computer, where both RAM and ROM chips are used:

A remote-controlled toy car has circuitry which contains both RAM and ROM chips. The remote control is a hand-held device. Explain the function of the RAM and ROM chip in this application.

We will consider the function of each type of memory independently:

ROM

- storing the factory settings such as remote control frequencies
- storing the 'start-up' routines when the toy car is first switched on
- storing of the set routines; for example, how the buttons on the hand-held device control turning left, acceleration, stopping, and so on.

RAM

- the user may wish to program in their own routines; these new instructions would be stored in the RAM chip
- the RAM chip will store the data/instructions received from the remote control unit.

Activity 3.6

1 Describe how ROM and RAM chips could be used in the following devices:
   - **a** a microwave oven
   - **b** a refrigerator
   - **c** a remote-controlled model aeroplane; the movement of the aeroplane is controlled by a hand-held device.

3.3.2 Secondary and off-line storage

Secondary (and off-line) storage includes storage devices that are not directly addressable by the CPU. They are non-volatile devices that allow data to be stored as long as required by the user. This type of storage can store more data than primary memory, but data access time is considerably longer than with RAM or ROM. All applications, the operating system, device drivers and general files (for example, documents, photos and music) are stored on secondary storage. The following section discusses the various types of secondary storage that can be found on the majority of computers.

3.3.3 Magnetic, optical and solid-state storage

Secondary (and off-line) storage falls into three categories according to the technology used:

- magnetic
- solid state
- optical.

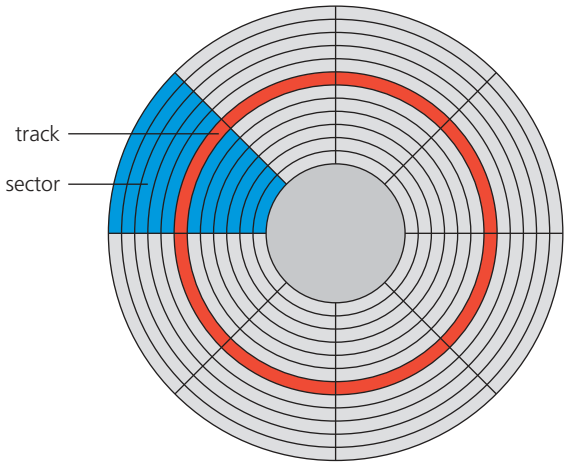
▲ **Figure 3.62** HDD

Magnetic storage

Hard Disk Drives (HDD)

Hard disk drives (HDD) are still one of the most common methods used to store data on a computer.

Data is stored in a digital format on the magnetic surfaces of the disks (or platters, as they are frequently called). The hard disk drive will have a number of platters that can spin at about 7000 times a second. Read-write heads consist of electromagnets that are used to read data from or write data to the platters. Platters can be made from aluminium, glass or a ceramic material. A number of read-write heads can access all of the surfaces of the platters in the disk drive. Normally each platter will have two surfaces which can be used to store data. These read-write heads can move very quickly – typically they can move from the centre of the disk to the edge of the disk (and back again) 50 times a second.



▲ **Figure 3.63** Tracks and sectors

Data is stored on the surface in sectors and tracks. A sector on a given track will contain a fixed number of bytes. Unfortunately, hard disk drives have very slow data access when compared to, for example, RAM. Many applications require the read-write heads to constantly look for the correct blocks of data; this means a large number of head movements. The effects of **latency** then become very significant. Latency is defined as the time it takes for a specific block of data on a data track to rotate around to the read-write head.

Users will sometimes notice the effect of latency when they see messages such as 'Please wait' or, at its worst, 'not responding'.

When a file or data is stored on a HDD, the required number of sectors needed to store the data will be allocated. However, the sectors allocated may not be adjacent to each other. Through time, the HDD will undergo numerous deletions and editing which leads to sectors becoming increasingly **fragmented** resulting in a gradual deterioration of the HDD performance (in other words, it takes longer and longer to access data). Defragmentation software can improve on this situation by 'tidying up' the disk sectors.

Link

See Section 4.1.1 for more on defragmentation.

All data in a given sector on a HDD will be read in order (that is, sequentially); however, access to the sector itself will be by a direct read/write head movement.

Removable hard disk drives are essentially HDDs external to the computer that can be connected to the computer using one of the USB ports. In this way, they can be used as a back-up device or another way of transferring files between computers.

Solid state drives (SSD)

Latency is an issue in HDDs as described earlier. Solid state drives (SSD) remove this issue considerably since they have no moving parts and all data is retrieved at the same rate. They don't rely on magnetic properties; the most common type of solid state storage devices store data by controlling the movement of electrons within NAND or NOR chips. The data is stored as 0s and 1s in millions of tiny transistors (at each junction one transistor is called a floating gate and the other is called a control gate) within the chip. This effectively produces a non-volatile rewritable memory.

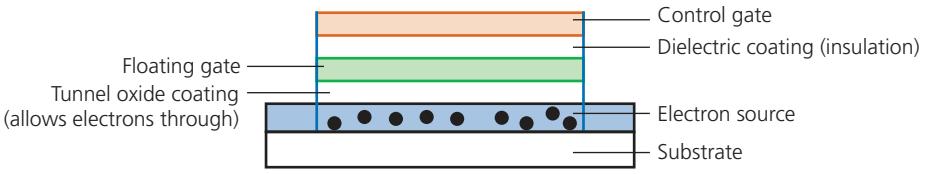
Link

For more on NAND and NOR gates see Chapter 10.

Floating gate and control gate transistors

Floating gate and control gate transistors use CMOS (complementary metal oxide semi-conductor) NAND technology. Flash memories make use of a matrix; at each

intersection on the matrix there is a floating gate and a control gate arranged as follows:



▲ **Figure 3.64** Flash memory

A dielectric coating separates the two transistors, which allows the floating gate transistor to retain its charge (which is why the memory is non-volatile). The floating gate transistor has a value of 1 when it is charged and a value of 0 when it isn't. To program one of these 'intersection cells' a voltage is applied to the control gate and electrons from the electron source are attracted to it. But due to the dielectric coating, the electrons become trapped in the floating gate. Hence, we have control over the bit value stored at each intersection. (**Note:** After about 12 months, this charge can leak away, which is why a solid state device should be used at least once a year to be certain it will retain its memory.)

The main benefits of this newer solid state technology over hard disk drives are:

- they are more reliable (no moving parts to go wrong)
- they are considerably lighter (which makes them suitable for laptops)
- they don't have to 'get up to speed' before they work properly
- they have a lower power consumption
- they run much cooler than HDDs (both these points again make them very suitable for laptop computers)
- because of no moving parts, they are very thin
- data access is considerably faster than HDD.

The main drawback of SSD is still the longevity of the technology (although this is becoming less of an issue). Most solid state storage devices are conservatively rated at only 20 GB of write operations per day over a three year period – this is known as **SSD endurance**. For this reason, SSD technology is still not used in all servers, for example, where a huge number of write operations take place every day. However, the durability of these solid state systems is being improved by a number of manufacturers and they are rapidly becoming more common in applications such as servers and **cloud storage** devices.

Note: It is also not possible to over-write existing data on a flash memory device; it is necessary to first erase the old data and then write the new data at the same location.

Memory sticks/flash memories

Memory sticks/flash memories (also known as pen drives) use solid state technology.

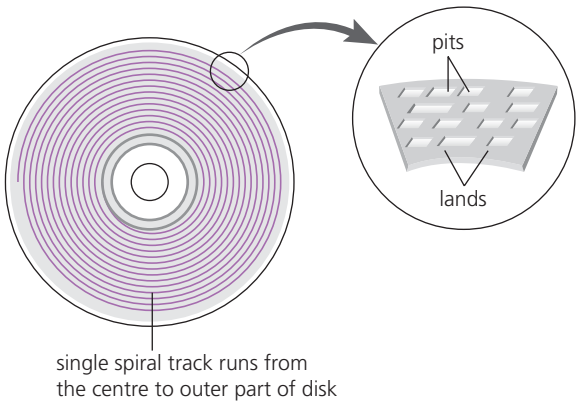
They usually connect to the computer through the USB port. Their main advantage is that they are very small, lightweight devices, which make them very suitable as a method for transferring files between computers. They can also be used as small back-up devices for music or photo files, for example.

Complex or expensive software, such as financial planning software, often uses memory sticks as a dongle. The dongle contains additional files that are needed to run the software. Without this dongle, the software won't work properly. It therefore prevents illegal or unauthorised use of the software, and also prevents copying of the software since, without the dongle, it is useless.

Optical media

CD/DVD disks

CDs and **DVDs** are described as **optical storage devices**. Laser light is used to read and write data to and from the surface of the disk.

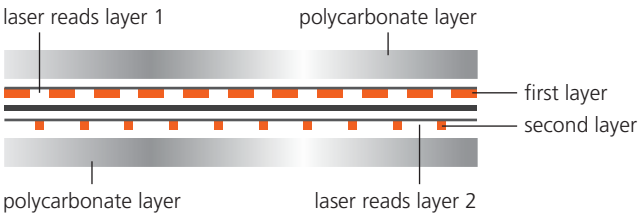


▲ **Figure 3.65** Optical media

Both CDs and DVDs use a thin layer of metal alloy or light-sensitive organic dye to store the data. As can be seen from the diagram in Figure 3.65, both systems use a single, spiral track which runs from the centre of the disk to the edge. When a disk spins, the optical head moves to the point where the laser beam 'contacts' the disk surface and follows the spiral track from the centre outwards. As with a HDD, a CD/DVD is divided into sectors allowing direct access to data. Also, as in the case of HDD, the outer part of the disk runs faster than the inner part of the disk.

The data is stored in 'pits' and 'lands' on the spiral track. A red laser is used to read and write the data. CDs and DVDs can be designated 'R' (write once only) or 'RW' (can be written to or read from many times).

DVD technology is slightly different to that used in CDs. One of the main differences is the potential for **dual-layering**, which considerably increases the storage capacity. Basically, this means that there are two individual recording layers. Two layers of a standard DVD are joined together with a transparent (polycarbonate) spacer, and a very thin reflector is also sandwiched between the two layers. Reading and writing of the second layer is done by a red laser focusing at a fraction of a millimetre difference compared to the first layer.



▲ **Figure 3.66** Dual-layering on a DVD

Standard, single layer DVDs still have a larger storage capacity than CDs because the 'pit' size and track width are both smaller. This means that more data can be stored on the DVD surface. DVDs use lasers with a wavelength of 650 nanometres; CDs use lasers with a wavelength of 780 nanometres. The shorter the wavelength of the laser light, the greater the storage capacity of the medium.

Blu-ray discs

Blu-ray discs are another example of optical storage media. However, they are fundamentally different to DVDs in their construction and in the way they carry out read-write operations.

Note: it is probably worth mentioning why they are called Blu-ray rather than Blue-ray; the simple reason is it was impossible to copyright the word 'Blue' and hence the use of the word 'Blu'.

The main differences between DVD and Blu-ray are:

▲ **Figure 3.67** Blu-ray disc

- a blue laser, rather than a red laser, is used to carry out read and write operations; the wavelength of blue light is only 405 nanometres (compared to 650 nm for red light)
- using blue laser light means that the 'pits' and 'lands' can be much smaller; consequently, Blu-ray can store up to five times more data than normal DVD
- single-layer Blu-ray discs use a 1.2 mm thick polycarbonate disk; however, dual-layer Blu-ray and normal DVDs both use a sandwich of two 0.6 mm thick disks (i.e. 1.2 mm thick)
- Blu-ray disks automatically come with a secure encryption system that helps to prevent piracy and copyright infringement
- the data transfer rate for a DVD is 10 Mbps and for a Blu-ray disc it is 36 Mbps (this equates to 1.5 hours to transfer 25 GiB of data).

Since Blu-ray discs can come in single layer or dual-layer format (unlike DVD, which is always dual-layer), it is probably worth also comparing the differences in capacity and interactivity of the two technologies.

Comparison of the capacity and interactivity of DVDs and Blu-ray discs

- A standard dual-layer DVD has a storage capacity of 4.7 GB (enough to store a 2-hour standard definition movie)
- A single-layer Blu-ray disc has a storage capacity of 27 GB (enough to store a 2-hour high definition movie or 13 hours of standard definition movies)
- A dual-layer Blu-ray disc has a storage capacity of 50 GB (enough to store 4.5 hours of high definition movies or 20 hours of standard definition movies).

Blu-ray allows greater interactivity than DVDs. For example, with Blu-ray, it is possible to:

- record high definition television programs
- skip quickly to any part of the disc
- create playlists of recorded movies and television programmes
- edit or re-order programmes recorded on the disc
- automatically search for empty space on the disc to avoid over-recording
- access websites and download subtitles and other interesting features.

Finally, Table 3.11 summarises the main differences between CDs, DVDs and Blu-ray.

All these optical storage media are used as back-up systems (for photos, music and multimedia files). This also means that CDs and DVDs can be used to transfer files between computers. Manufacturers sometimes supply their software (e.g. printer drivers) using CDs and DVDs. When the software is supplied in this way, the disk is usually in a read-only format.

▼ **Table 3.11** Comparison of CD, DVD and Blu-ray (Note: nm = 10^{-9} metres and µm = 10^{-6} metres)

Disk type	Laser colour	Wavelength of laser light	Disk construction	Track pitch (distance between tracks)
CD	Red	780 nm	single 1.2 mm polycarbonate layer	1.60 µm
DVD (dual-layer)	Red	650 nm	two 0.6 mm polycarbonate layers	0.74 µm
Blu-ray (single layer)	Blue	405 mm	single 1.2 mm polycarbonate layer	0.30 µm
Blu-ray (dual-layer)	Blue	405 nm	two 0.6 mm polycarbonate layers	0.30 µm

The most common use of DVD and Blu-ray is the supply of movies or games. The memory capacity of CDs isn't big enough to store most movies (see earlier comparison notes).

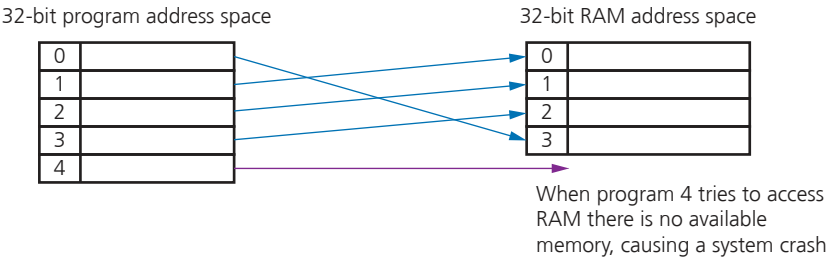
3.3.4 Virtual memory

One of the problems associated with memory management is the case when processes run out of RAM. If the amount of available RAM is exceeded due to multiple programs running, it is likely to cause a system crash. This can be solved by utilising the hard disk drive (or SSD) if we need more memory. This is the basis behind **virtual memory**. Essentially RAM is the ***physical memory***, while ***virtual memory*** is RAM + swap space on the hard disk or SSD.

To execute a program, data is loaded into memory from HDD (or SSD) whenever required. It is possible to show the differences between using normal memory management and virtual memory management in two simple diagrams.

Without virtual memory

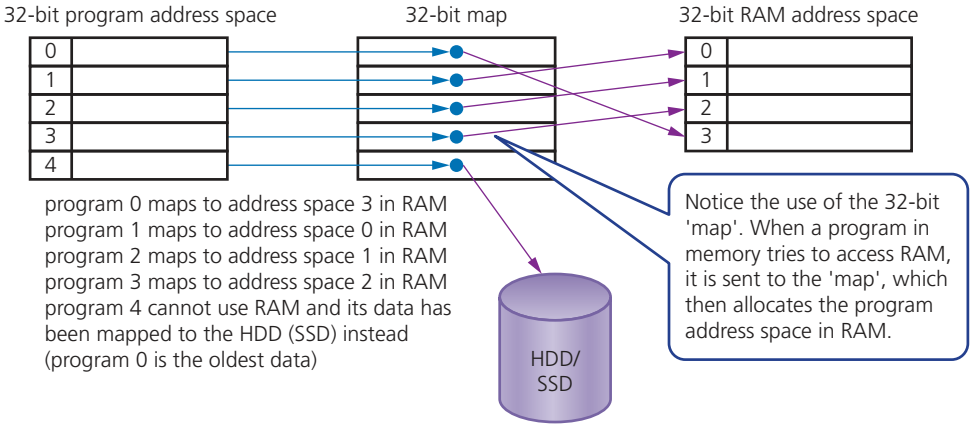
Suppose we have five programs (numbered 0 to 4) that are in memory, all requiring access to RAM. The first diagram shows what would happen without virtual memory being used (i.e. the computer would run out of RAM memory space):



▲ **Figure 3.68** Normal memory management

With virtual memory

We will now consider what happens if the CPU uses virtual memory to allow all five programs to access RAM as required. This will require moving data out of RAM into HDD/SSD and then allowing other data to be moved out of HDD/SSD into RAM:

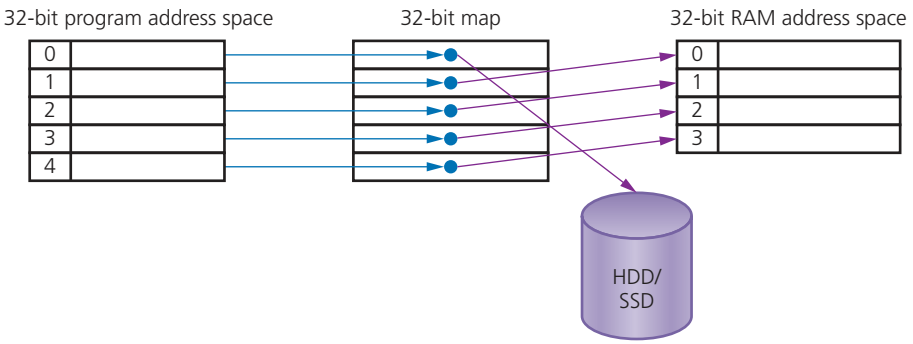


▲ **Figure 3.69** Status just before program 4 is given RAM space

Virtual memory now moves the oldest data out of RAM into the HDD/SSD to allow program 4 to gain access to RAM. The 32-bit 'map' is now updated to reflect this new situation:

- data from program 0 (which was using RAM address space 3 – the oldest data) is now mapped to the HDD/SSD instead, leaving address space 3 free for use by program 4
- program 4 now maps to address space 3 in RAM, which means program 4 now has access to RAM.

Our diagram now changes to:



▲ **Figure 3.70** Status with program 0 now mapped to HDD and program 4 has access to RAM

All of this will continue to occur until RAM is no longer being over-utilised by the competing programs running in memory. Virtual memory gives the illusion of unlimited memory being available. Even though RAM is full, data can be moved in and out of the HDD/SSD to give the illusion that there is still memory available. In computer operating systems, **paging** is used by memory management to store and retrieve data from HDD/SSD and copy it into RAM. A **page** is a fixed-length consecutive (or contiguous) block of data utilised in virtual memory systems.

This is a key part of how virtual memory works allowing **data blocks** (pages) to be moved in and out of a HDD/SSD. However, accessing data in virtual memory is slower so, as mentioned earlier on in this chapter, the larger the RAM the faster the CPU can operate. This is one of the benefits of increasing RAM size as far as possible.

The main benefits of virtual memory are:

- programs can be larger than physical memory and still be executed
- there is no need to waste memory with data that isn't being used (e.g. during error handling)
- it reduces the need to buy and install more expensive RAM memory (although as mentioned earlier there are limits to the value of doing this).

When using HDD for virtual memory the main drawback is **disk thrashing**. As main memory fills, more and more data needs to be swapped in and out of virtual memory leading to a very high rate of hard disk read/write head movements; this is known as disk thrashing. If more and more time is spent on moving data in and out of memory than actually doing any processing, then the processing speed of the computer will be considerably reduced. A point can be reached when the execution of a process comes to a halt since the system is so busy moving data in and out of memory rather than doing any actual execution – this is known as the **thrash point**. Due to large numbers of head movements, this can also lead to premature failure of a hard disk drive. Thrashing can be reduced by installing more RAM, reducing the number of programs running at a time or reducing the size of the swap file. Another way of reducing this problem is to make use of a solid state drive (SSD) rather than using HDD.

3.3.5 Cloud storage

Public and private cloud computing

Cloud storage is a method of data storage where data is stored on remote servers. The same data is stored on more than one server in case of maintenance or repair, allowing clients to access data at any time. This is known as **data redundancy**. The physical environment is owned and managed by a hosting company and may include hundreds of servers in many locations.

There are three common systems:

- Public cloud – this is a storage environment where the customer/client and cloud storage provider are different companies
- Private cloud – this is storage provided by a dedicated environment behind a company firewall; customer/client and cloud storage provider are integrated and operate as a single entity
- Hybrid cloud – this is a combination of the two above environments; some data resides in the private cloud and less sensitive/less commercial data can be accessed from a public cloud storage provider.

Instead of saving data on a local hard disk or other storage device, a user can save their data 'in the cloud'. The benefits and drawbacks of using cloud storage are shown in Table 3.12.

▼ **Table 3.12** Benefits and drawbacks of cloud storage

Benefits of using cloud storage	Drawbacks of using cloud storage
customer/client files stored on the cloud can be accessed at any time from any device anywhere in the world provided internet access is available	if the customer/client has a slow or unstable internet connection, they would have many problems accessing or downloading their data/files
there is no need for a customer/client to carry an external storage device with them, or even use the same computer to store and retrieve information	costs can be high if large storage capacity is required; it can also be expensive to pay for high download/upload data transfer limits with the customer/client internet service provider (ISP)
the cloud provides the user with remote back-up of data with obvious benefits to alleviate data loss/disaster recovery	the potential failure of the cloud storage company is always possible – this poses a risk of loss of all back-up data
if a customer/client has a failure of their hard disk or back-up device, cloud storage will allow recovery of their data	
the cloud system offers almost unlimited storage capacity	

Data security when using cloud storage

Companies that transfer vast amounts of confidential data from their own systems to a cloud service provider are effectively relinquishing control of their own data security. This raises a number of questions:

- what physical security exists regarding the building where the data is housed?
- how good is the cloud service provider's resistance to natural disasters or power cuts?
- what safeguards exist regarding personnel who work for the cloud service company; can they use their authorisation codes to access confidential data for monetary purposes?

Potential data loss when using cloud storage

There is a risk that important and irreplaceable data could be lost from the cloud storage facilities. Actions from hackers (gaining access to accounts or pharming attacks, for example) could lead to loss or corruption of data. Users need to be certain that sufficient safeguards exist to overcome these risks.

The following breaches of security involving some of the largest cloud service providers suggest why some people are nervous of using cloud storage for important files:

- The XEN security threat, which forced several cloud operators to reboot all their cloud servers, was caused by a problem in the XEN hypervisor (a hypervisor is a piece of computer software, firmware or hardware that creates and runs virtual machines).
- A large cloud service provider permanently lost data during a routine back-up procedure.
- The celebrity photos cloud hacking scandal, in which more than 100 private photos of celebrities were leaked. Hackers had gained access to a number of cloud accounts, which enabled them to publish the photos on social networks and sell them to publishing companies.
- In 2016, the National Electoral Institute of Mexico suffered a cloud security breach in which 93 million voter registrations, stored on a central database, were compromised and became publicly available to everyone. To make matters worse, much of the information on this database also linked to a cloud server outside Mexico.

Activity 3.7

1 Name two types of memory used in a mobile phone. For each named memory, describe its purpose in the mobile phone.

2 **a** Explain what is meant by ***virtual memory***.

b Five programs are currently being run in a computer. Program 1 is using 10 GiB of RAM, program 2 is using 5 GiB of RAM, program 3 is using 12 GiB of RAM and program 4 is using 4 GiB of RAM. The programs are at the stage where program 5 now needs to access RAM, but RAM is presently full (RAM has a 32 GiB maximum capacity). Explain how virtual memory could be used to allow program 5 to access RAM without any of the data from the other four programs being lost.

3 Five descriptions of computer terms are shown on the left and five terms are shown on the right. Draw lines to connect each description to the correct computer term.

storage environment where the client and remote storage provider are different companies	thrashing
high rate of HDD read/write operations causing large amount of head movement	swap space
space on HDD or SSD reserved for data used in virtual memory management	cloud storage
situation where a HDD is so busy doing read/write operations that execution of a process is halted	thrash point
method of data storage where the data is stored on hundreds of off-site servers	public cloud

4 **a** Give four differences between RAM and ROM chips.

b Give an example of the use of each type of memory.

5 **a** Use the following words to complete the paragraph below which describes how solid state memories work (each word may be used once, more than once or not at all).

Word list:		
	control gate	NAND
	electrons	negative
	floating gate	positive
	intersection	transistor
	matrix	volatile

Solid state devices control the movement of ..... within a ..... chip. The device is made up of a ..... and at each ..... there is a ..... and a ..... transistor. ..... are attracted towards ..... when a voltage is applied.

b Give three advantages of using SSD, rather than HDD, which make SSD technology particularly suitable for use in laptop computers.

c Describe one disadvantage of solid state technology.

3.4 Network hardware

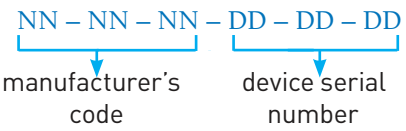
3.4.1 Network interface card (NIC)

A **network interface card (NIC)** is needed to allow a device to connect to a network (such as the internet). It is usually part of the device hardware and contains the **Media Access Control (MAC)** address generated at the manufacturing stage.

Wireless network interface cards/controllers (WNICs) are the same as NICs in that they are used to connect devices to the internet or other networks. However, they use wireless connectivity utilising an antenna to communicate with networks via microwaves. They would normally plug into the USB port or be part of an internal integrated circuit.

3.4.2 Media Access Control (MAC)

A MAC address is made up of 48 bits which are shown as six groups of hexadecimal digits with the general format:



For example, 00 – 1C – B3 – 4F – 25 – FF where the first six hex digits identify the device as made by, for example, Apple and the second set of six hex digits are the serial number of the device itself (this is unique). If the NIC card is replaced, the MAC address will also change.

Types of MAC address

It should finally be pointed out that there are two types of MAC address: the **Universally Administered MAC Address (UAA)** and the **Locally Administered MAC Address (LAA)**.

The UAA is by far the most common type of MAC address and this is the one set by the manufacturer at the factory. It is rare for a user to want to change this MAC address.

However, there are some occasions when a user or organisation wishes to change their MAC address. This is a relatively easy task to carry out, but it will cause big problems if the changed address isn't unique.

There are a few reasons why the MAC address needs to be changed using LAA:

- certain software used on mainframe systems need all the MAC addresses of devices to fall into a strict format; because of this, it may be necessary to change the MAC address of some devices to ensure they follow the correct format
- it may be necessary to bypass a MAC address filter on a router or a firewall; only MAC addresses with a certain format are allowed through, otherwise the devices will be blocked if their MAC address doesn't adhere to the correct format
- to get past certain types of network restrictions it may be necessary to emulate unrestricted MAC addresses; hence it may require the MAC address to be changed on certain devices connected to the network.

3.4.3 Internet protocol (IP) address

When a device connects to a private network, a router assigns a private IP address to it. That IP address is unique on that network, but might be the same as an IP address on a separate network. However, when a router connects to the internet it is given a unique public IP address. This is usually supplied by the internet service provider (ISP). No other device on the internet has the same public IP address. All the devices connected to that router have the same public IP address as the router but each have their own different private IP addresses on that network. Because the operation of the internet is based on a set of protocols (rules), it is necessary to supply an IP address. Protocols define the rules that must be agreed by senders and receivers of data communicating through the internet.

There are two versions of IP: IPv4 and IPv6. IPv4 is based on 32 bits and the address is written as four groups of eight bits (shown in denary format); for example,

254.25.28.77

Because the use of only 32 bits considerably reduces the potential number of devices and routers used on the internet at any one time, a newer version called IPv6 is now used. This uses 128-bit addresses that take the form of eight groups of hex digits; for example,

A8FB:7A88:FFF0:0FFF:3D21:2085:66FB:F0FA

Link

For more on packet routes see Section 2.1.1.

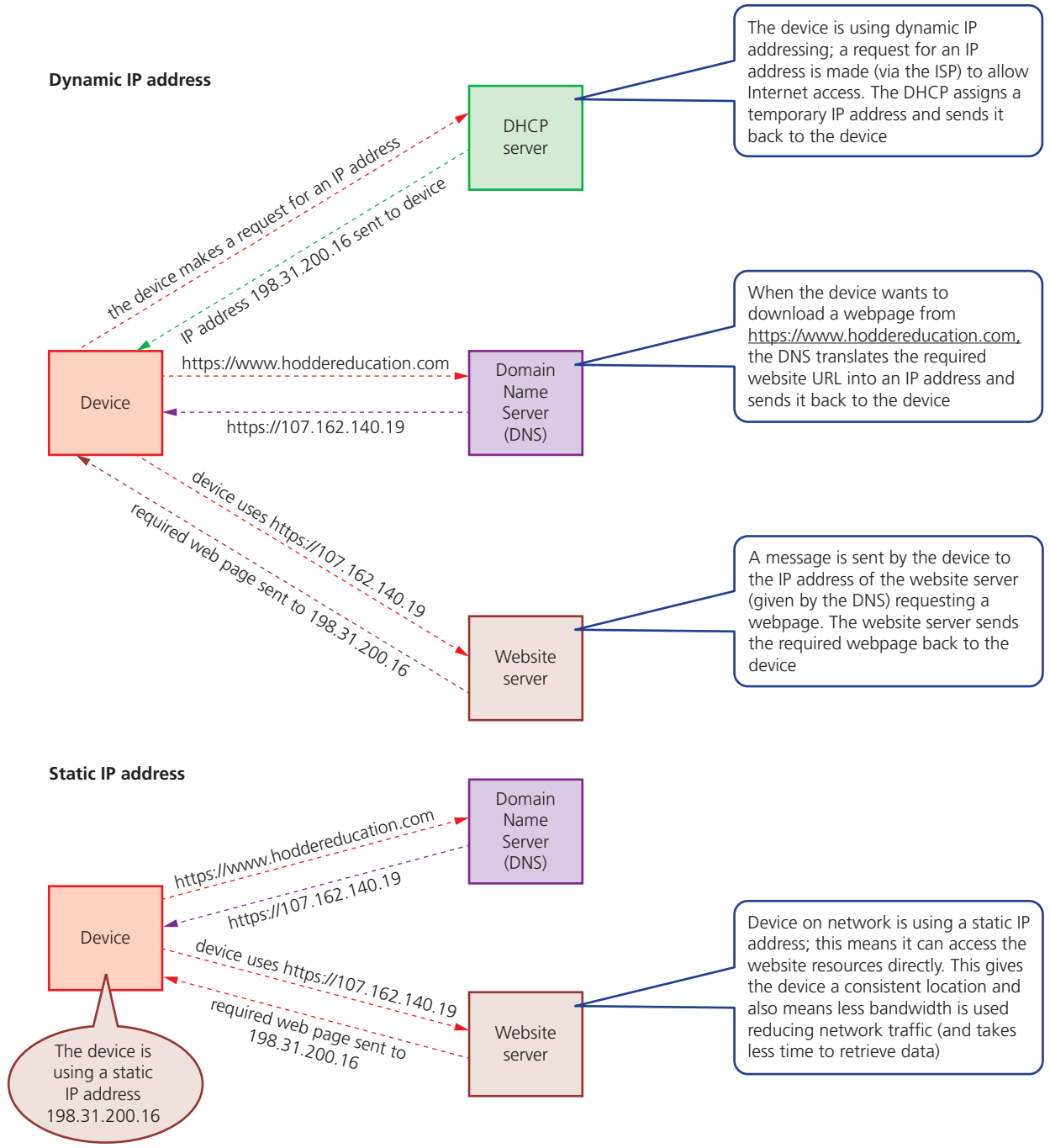
Note the use of colons (:) and hexadecimal numbering. IPv6 has been designed to allow the internet to grow in terms of the number of hosts and potential increase in the amount of data traffic. The main advantages of IPv6 compared to IPv4 are:

- removes the risk of IP address collisions
- has built-in authentication checks
- allows for more efficient packet routes.

Table 3.13 compares the features of MAC addresses and IP addresses:

▼ **Table 3.13** MAC addresses and IP addresses

MAC addresses	IP addresses
identifies the physical address of a device on the network	identifies the global address on the internet
unique for device on the network	may not necessarily be unique
assigned by the manufacturer of the device and is part of the NIC	dynamic IP addresses are assigned by ISP using DHCP each time the device connects to the internet (see later)
they can be universal or local	dynamic IP addresses change every time a device connects to the internet; static IP addresses don't change
when a packet of data is sent and received, the MAC address is used to identify the sender's and recipient's devices	used in routing operations as they specifically identify where the device is connected to the internet
use 48 bits	use either 32 bits (IPv4) or 128 bits (IPv6)
can be UAA or LAA	can be static or dynamic



▲ **Figure 3.71** Comparison of dynamic and static IP addressing

Static and dynamic IP addresses

IP addresses can be either **static** (don't change) or **dynamic** (change every time a device connects to the internet).

Static

Static IP addresses are permanently assigned to a device by the internet service provider (ISP); they don't change each time a device logs onto the internet. Static IP addresses are usually assigned to:

- remote servers which are hosting a website
- an online database
- a File Transfer Protocol (FTP) server. FTP servers are used when files need to be transferred to various computers throughout the network.

Dynamic

Dynamic IP addresses are assigned by the ISP each time a device logs onto the internet. This is done using **Dynamic Host Configuration Protocol (DHCP)**. A computer on the internet, configured as a DHCP server, is used by the ISP to automatically assign an IP address to a device. As the name suggests, a dynamic IP address could be different every time a device connects to the internet.

Table 3.14 compares static and dynamic IP addresses:

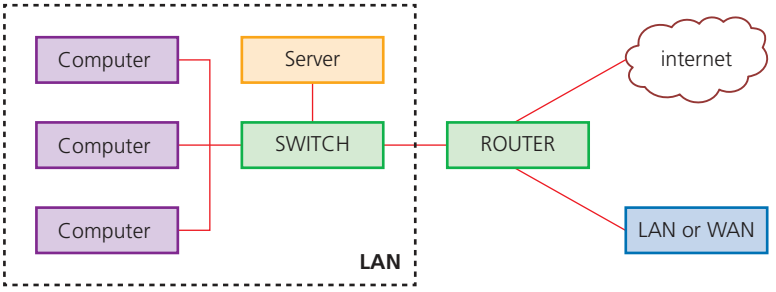
▼ **Table 3.14** Dynamic and static IP addresses

Dynamic IP addresses	Static IP addresses
greater privacy since they change each time a user logs on	since static IP addresses don't change, they allow each device to be fully traceable
dynamic IP addresses can be an issue when using, for example, VoIP since this type of addressing is less reliable as it can disconnect and change the IP address causing the VoIP connection to fail	allow for faster upload and download speeds (see Figure 3.71)
	more expensive to maintain since the device must be constantly running so that information is always available

Figure 3.71 shows the sequence of events when either a dynamic IP address or static IP address is assigned to a device using the internet. The diagram shows how a device contacts web servers that are also connected to the internet. A DHCP server supplies a dynamic IP address to the device, a DNS server looks up the domain name of the desired website into an IP address and a website server contains the web pages of the desired website.

3.4.4 Routers

Routers enable data packets to be routed between different networks, for example, to join a LAN to a WAN. The router takes data transmitted in one format from a network (which is using a particular protocol) and converts the data to a protocol and format understood by another network, thereby allowing them to communicate. A router would typically have an internet cable plugged into it and several cables connecting to computers and other devices on the LAN.



▲ **Figure 3.72** Router flow diagram

Broadband routers sit behind a firewall. The firewall protects the computers on a network. The router's main function is to transmit internet and transmission protocols between two networks and also allow private networks to be connected together.

Routers inspect the data package sent to it from any computer on any of the networks connected to it. Since every computer on the same network has the same part of an internet protocol (IP) address, the router is able to send the data packet to the appropriate switch, and the data will then be delivered to the correct device using the MAC destination address. If the MAC address doesn't match any device connected to the switch, it passes on to another switch on the same network until the appropriate device is found. Routers can be wired or wireless devices.

Activity 3.8

1 Explain each of the following terms:
   - **a** Network interface card (NIC)
   - **b** MAC address
   - **c** IP address
   - **d** Router
   - **e** DHCP server.

2 **a** Give two features of dynamic IP addresses.
   - **b** Give two features of static IP addresses.
   - **c** Explain why we need both types of IP address.

3 Describe three differences between MAC addresses and IP addresses.

Extension

For those students considering the study of this subject at A Level, the following section gives some insight into further study of computer hardware and wireless networks.

Topic 1: Computer ports

Input and output devices are connected to a computer via ports. The interaction of the ports with connected input and output is controlled by the control unit. Here we will summarise some of the more common types of ports found on modern computers:

USB ports

The **Universal Serial Bus (USB)** is an **asynchronous serial** data transmission method. It has quickly become the standard method for transferring data between a computer and a number of devices. Essentially, the USB cable consists of:

- a 4-wired shielded cable
- 2 of the wires are used for power and the earth
- 2 of the wires are used in the data transmission.

When a device is plugged into a computer, using one of the USB ports, the computer:

- automatically detects that a device is present (this is due to a small change in the voltage level on the data signal wires in the cable)
- the device is automatically recognised, and the appropriate device driver is loaded up so that computer and device can communicate effectively
- if a new device is detected, the computer will look for the device driver which matches the device; if this is not available, the user is prompted to download the appropriate software.

Even though the USB system has become the industrial standard, there are still a number of benefits (✓) and drawbacks (X) to using this system:

✓	X
devices plugged into the computer are automatically detected; device drivers are automatically loaded up	the present transmission rate is limited to less than 500 megabits per second
the connectors can only fit one way; this prevents incorrect connections being made	the maximum cable length is presently about 5 metres
USB has become the industry standard; this means that considerable support is available to users	the older USB standard (1.1) may not still be supported in the near future
several different data transmission rates are supported	
newer USB standards are backward compatible with older USB standards	

High-definition Multimedia Interface (HDMI)

High-definition Multimedia Interface (HDMI) ports allow output (both audio and visual) from a computer to an HDMI-enabled monitor or other device. It will support high definition signals (enhanced or standard). HDMI was introduced as a digital replacement for the older VGA analogue system. Modern HD (high definition) televisions have the following features which are making VGA a redundant technology:

- they use a widescreen format (16:9 aspect ratio)
- the screens use a greater number of pixels (typically 1920 ×1080)
- the screens have a faster refresh rate (e.g. 120 Hz or 120 frames a second)
- the range of colours is extremely large (some companies claim up to 4 million different colour variations!).

This all means that modern HD televisions require more data and this data has to be received at a much faster rate than with older televisions (e.g. 10 gigabits per second). HDMI increases the bandwidth making it possible to supply the necessary data to produce high quality sound and visual effects.

Topic 2: Wired and wireless networks

Wireless

Wi-Fi and Bluetooth

Both Wi-Fi and Bluetooth offer wireless communication between devices. They both use electromagnetic radiation as the carrier of data transmission.

Bluetooth sends and receives radio waves in a band of 79 different frequencies (known as channels). These are all centred on a 2.45 GHz frequency. Devices using Bluetooth automatically detect and connect to each other; but they don't interfere with other devices since each communicating pair uses a different channel (from the 79 options).

When a device wants to communicate, it picks one of the 79 channels at random. If the channel is already being used, it randomly picks another channel. This is known as spread-spectrum frequency hopping. To further minimise the risks of interference with other devices, the communication pairs constantly change the frequencies (channels) they are using (several times a second). Bluetooth creates a secure wireless personal area network (WPAN) based on key encryption.

Essentially, Bluetooth is useful:

- when transferring data between two or more devices which are very close together (<30 metres distance)
- when the speed of data transmission is not critical
- for low bandwidth applications (e.g. when sending music files from a mobile phone to a headset).

Wi-Fi is best suited to operating full-scale networks since it offers much faster data transfer rates, better range and better security than Bluetooth. A Wi-Fi-enabled device (such as a computer or smartphone) can access the internet wirelessly at any wireless access point (WAP) or 'hot spot' up to 100 metres away.

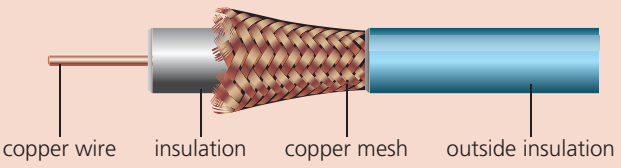
As mentioned above, wireless connectivity uses electromagnetic radiation: radio waves, microwaves or infrared. The scale of frequency and wavelength of magnetic radiation can be seen in the table below:

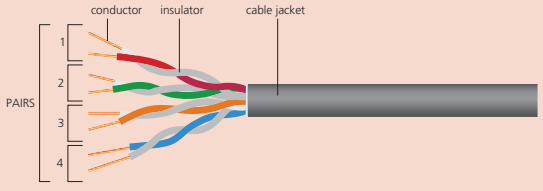
	Radio waves	Microwaves	Infrared	Visible light	Ultra violet	X-rays	Gamma rays
Wave length (m)	102	10^{-1}	10^{-5}	10^{-5}	10^{-7}	10^{-9}	10^{-11}
◄							
Frequency (Hz)	3 MHz	3 GHz	300 GHz	30 THz	3 PHz	300 PHz	30 EHz

Wired

There are three main types of cable used in wired networks:

- twisted pair
- coaxial
- fibre optic.





Twisted pair cable

Coaxial cable

Fibre optic cable

Wired versus wireless

When deciding whether a network should use wired or wireless connectivity, the following factors should be considered:

Wireless networking	Wired networking
it is easier to expand the networks and it isn't necessary to connect the devices using cables	using cables produces a more reliable and stable network; wireless connectivity is often subject to interference
this gives devices increased mobility provided they are within range of the WAPs	data transfer rates tend to be faster and there won't be any 'dead spots'
there is an increased chance of interference from external sources	setting up cabled networks tends to be cheaper overall in spite of the need to buy and install cable
data is less secure than with wired systems; it is easier to intercept radio waves and microwaves than cables; it is essential to protect data transmissions using encryption (e.g. WEP, WPA2)	however, cabled networks lose the ability for devices to be mobile; they must be close enough to allow for cable connections
the data transmission rate is still slower than for cabled networks although it continues to improve	having lots of wires can lead to a number of hazards such as tripping hazards, overheating of connections (leading to potential fire risk) and disconnection of cables during routine office cleaning
it is possible for signals to be stopped by thick walls (e.g. in old houses) and there may be areas of variable signal strength leading to 'drop out'	

In this chapter, you have learnt about:

- ✔ the role of a CPU/microprocessor
- ✔ von Neumann architecture
- ✔ the Fetch–Decode–Execute cycle
- ✔ the function of a cache, (system) clock and multi-core core CPUs
- ✔ instruction sets
- ✔ embedded systems
- ✔ the operation of a number of input devices
- ✔ the operation of a number of output devices
- ✔ the use of sensors in a number of control and monitoring applications
- ✔ primary storage
- ✔ secondary storage (magnetic, optical and solid state)
- ✔ virtual memories
- ✔ cloud storage
- ✔ network interface card (NIC)
- ✔ MAC and IP addressing
- ✔ routers in networks.

Key terms used throughout this chapter

central processing unit (CPU) – responsible for the execution or processing of all the instructions and data in a computer

integrated circuit – usually a chip made from a semi-conductor material which carries out the same tasks as a larger circuit made from individual components

von Neumann architecture – a type of computer architecture which introduced the concept of the stored program in the 1940s

Arithmetic & Logic Unit (ALU) – the component of the CPU that carries out all arithmetic and logical operations

accumulator (ACC) – temporary general-purpose register that stores numerical values at any part of a given operation

memory address register (MAR) – a register that stores the address of the memory location currently being read from or written to

current instruction register (CIR) – a register that stores the current instruction being decoded and executed

memory data register (MDR) – a register that stores data that has just been read from memory or data that is about to be written to memory

program counter (PC) – a register that stores the address where the next instruction to be read can be found

control unit – the component of a computer's CPU that ensures synchronisation of data flow and programs throughout the computer by sending out control signals along the control bus

system clock – produces timing signals on the control bus to ensure synchronisation takes place

clock cycle – clock speeds are measured in terms of GHz; this is the vibrational frequency of the system clock which sends out pulses along the control bus; for example, a 3.5 GHZ clock cycle means 3.5 billion clock cycles a second

immediate access store (IAS) – memory that holds all data and programs needed to be accessed by the control unit

backing store – a secondary storage device (such as HDD or SSD) used to store data permanently even when the computer is powered down

cache – is temporary memory using static RAM to hold frequently used data/instructions by the CPU thereby increasing CPU performance. More generally, cache means any area of storage used to quickly access frequently-used data - other examples include web cache, database cache, DNS cache

register – a temporary component in the CPU which can be general or specific in its use; it holds data or instructions as part of the Fetch–Decode–Execute cycle

address – a label for a memory location used by the CPU to track data

memory location – a numbered place in memory where values can be stored

system buses – a connection between major components in a computer that can carry data, addresses or control signals

address bus – the system bus that carries the addresses throughout the computer system

data bus – the system bus that allows data to be carried from CPU to memory (and vice versa) or to and from input/output devices

control bus – the system bus that carries signals from control unit to all other computer components

unidirectional – can travel in one direction only; used to describe data

bidirectional – can travel in both directions; used to describe data

word – a group of bits used by a computer to represent a single unit; for example, modern computers often use 64-bit word lengths

overclocking – changing the clock speed of a system clock to a value higher than the factory/recommended setting

core – a unit on a CPU made up of an ALU, control unit and registers; a CPU may contain a number of cores

dual core – a CPU containing two cores

quad core – a CPU containing four cores

Fetch–Execute–Decode – a cycle in which instructions and data are fetched from memory, decoded and finally executed

Basic Input/Output System (BIOS) – a suite of programs on firmware that are used to perform the initialisation of a computer system during the boot-up process

opcode – part of a machine code instruction that identifies what action the CPU has to perform

operand – part of a machine code instruction that identifies what data is to be used

instruction set – the complete set of machine code instructions used a particular microprocessor

embedded system – a combination of hardware and software designed to carry out a specific set of functions

barcode – a series of dark and light lines of varying thickness used to represent data; the code has to be scanned using laser or LED light source

key field – the field that uniquely identifies a record in a file

quick response (QR) code – a matrix of dark and light squares which represent data; the pattern can be read and interpreted using a smartphone camera and QR app

frame QR code – a type of QR code that includes a space for advertising

DAC (digital to analogue converter) – device that converts digital data into electric currents that can drive motors, actuators and relays, for example

ADC (analogue to digital converter) – a device that converts analogue data (for example, data read from sensors) into a form understood by a computer

charge couple device (CCD) – a light sensitive cell made up of millions of tiny sensors acting as photodiodes

virtual keyboard – an onscreen keyboard which uses the features of the touch screen to emulate a physical keyboard

touch screen – a screen that allows the user to select or manipulate a screen image using the touch of a finger or stylus; touch screens most frequently use capacitive, infra-red or resistive technology

repetitive strain injury (RSI) – pain felt in the muscles, nerves and tendons caused by a repetitive action (for example, excessive clicking of a mouse button over a period of time)

optical mouse – a pointing device that uses a red LED to track the movement of the device and then relays its coordinates to a computer

pointing device – an input device that allows the user to control the movement of an onscreen cursor or to allow onscreen selection by clicking a button on the device

complementary metal oxide semi-conductor (CMOS) – a chip that generates an electric current (or pulses) when light falls on its surface

digital signal processor (DSP) – a processor that calculates, for example, the coordinates of a pointing device based on the pulses of electricity received

optical character recognition – technology that can convert hard copy text or images into a digital format to be stored in a computer memory

computer sided design (CAD) – software used to create drawings (for example, to send to a 3D printer or to produce blue-prints of a microprocessor design)

computed tomographic (CT) scanner – technology that can create a 3D image of a solid object by slicing up the object into thin layers (tomography)

capacitive touch screen – a type of touch screen that uses the change in the screen's capacitance (the ability to store an electrical charge) when it is touched by a finger or stylus

infra-red touch screen – a type of touch screen that uses infra-red beams and sensors to detect where the screen has been touched

resistive touch screen – a type of touch screen that uses two conductive layers which make contact where the screen has been touched

actuator – an output device that converts electrical energy into mechanical movement

digital micromirror device (DMD) – a chip that uses millions of tiny mirrors on its surface to create a video display

thermal bubble – inkjet printer technology whereby tiny resistors create heat and form an ink bubble which is ejected onto paper in an inkjet printer

piezoelectric crystal – a crystal located in an ink reservoir within an inkjet printer; the crystal vibrates and forces ink out onto paper

direct 3D printing – a 3D printing technique in which the print head moves in the x, y and z directions

binder 3D printing – a 3D printing method that uses a two-stage pass; the first stage uses dry powder and second stage uses a binding agent

cathode – a negative electrode

snode – a positive electrode

organic LED (OLED) – a light-emitting diode that uses the movement of electrons between a cathode and an anode to produce an on-screen image; it generates its own light so no backlighting is required

loudspeaker – an output device that converts electric current into sound

memory – the devices within the computer that are directly accessible by the CPU; there are two types of memory – RAM and ROM; memory is different to hard disk drives, for example, which are known as storage devices

random access memory (RAM) – primary memory that can be written to or read from

read only memory (ROM) – primary memory that cannot be written to (changed) and can only be read

dynamic RAM (DRAM) – a type of RAM chip that needs to be constantly refreshed

static RAM (SRAM) – a type of RAM chip that uses flip flops and doesn't need to be constantly refreshed

volatile – describes memory that loses its contents when the power is turned off

refresh – recharge every few seconds in order to maintain charge; for example with a device such as a capacitor

flip flop – electronic circuit with only two stable conditions

latency – the lag in a system; for example, the time it takes to find a track on a hard disk, which depends on the time it takes for the disk to rotate around to its read-write head

SSD endurance – the total guaranteed number of times data can be written to or read from a solid state drive (SSD) in its usable life cycle

optical storage – a type of storage that uses laser light to read and write data, and includes CDs, DVDs and Blu-ray discs

dual layering – using two recording layers in storage media such as DVDs and some Blu-rays

virtual memory – a memory management system that makes use of secondary storage and software to enable a computer to compensate for the shortage of actual physical RAM memory

disk thrashing (HDD) – a problem in a hard disk drive (HDD) caused by excessive swapping in and out of data causing a high rate of head movements during virtual memory operations

thrash point – the point at which the execution of a program comes to a halt because the system is so busy moving data in and out of memory rather than actually executing the program

data redundancy – the unnecessary storing of the same data on several storage devices at the same time

cloud storage – a method of data storage where data is stored on offsite servers; the physical storage may be on hundreds of servers in many locations

network interface card (NIC) – a hardware component (circuit board or chip) that is required to allow a device to connect to a network, such as the internet

router – a device that enables data packets to be moved between different networks, for example, to join a LAN to a WAN

static IP address – an IP address that doesn't change

MAC address – a unique identifier which acts as a network address for a device; it takes the form NN-NN-NN-DD-DD-DD, where NN is the manufacturer code and DD is the device code

dynamic IP address – a temporary IP address assigned to a device each time it logs onto a network

dynamic host configuration protocol (DHCP) – a server that automatically provides and assigns an IP address

Exam-style questions

1 a Many mobile phone and tablet manufacturers are moving to OLED screen technology. Give **three** reasons why this is happening. [3]

b A television manufacturer makes the following advertising claim:
"Our OLED screens allow the user to enjoy over 1 million vivid colours in true-to-life vision"
Comment on the validity of this claim by the manufacturer. [4]

2 a A company is developing a new games console. The console games will be stored on a ROM chip once the program to run the new game has been fully tested and developed.

i Give **two** advantages of putting the game's program on a ROM chip.

ii The manufacturers are also using RAM chips on the internal circuit board. Why have they done this?

iii The games console will have four USB ports. Apart from the need to attach games controllers, give reasons why USB ports are incorporated. [8]

b During development of the games console the plastic parts are being made by a 3D printer. Give **two** reasons why the manufacturer would use 3D printers. [2]

3 An air conditioning unit in a car is being controlled by a microprocessor and a number of sensors.

a Describe the main differences between **control** and **monitoring** of a process. [2]

b Describe how the sensors and microprocessor would be used to control the air conditioning unit in the car. Name at least **two** different sensors that might be used and explain the role of positive feedback in your description.
You might find drawing a diagram of your intended process to be helpful. [6]

4 a Describe the differences between a static IP address and a dynamic IP address. Include in your explanation, why both types of IP addressing are used. [4]

b What is meant by a MAC address? Describe the **two** different types of MAC address. [4]

5 a Five statements about two types of RAM memory (DRAM and SRAM) are shown below.
By drawing lines, link each statement to the correct type of RAM.

the data has to be refreshed constantly in order to retain data	DRAM
this type has the more complex circuitry	
it does not need to be refreshed as long as the power supply is still on	SRAM
it requires higher power consumption which is significant when installed in battery-powered devices, such as a laptop	
it is mostly used in the cache memory of the CPU where operational speed is important	

[5]

b Describe **three** of the differences between RAM and ROM. [3]

c Compare the two optical storage devices: DVD and Blu-ray. Your answer should include:
- technology differences
- capacity differences
- other features of the two types of storage. [6]

6 a Name **two** types of touchscreen technology used on mobile phones. For each named technology, describe how the position of where a finger touched the screen can be identified. [6]

b Organic LED (OLED) screens are now becoming increasingly common.
- **i** Briefly describe the technology behind OLED screens.
- **ii** Give **three** advantages of OLED screens when compared to LED/LCD screens. [3]

7 A zoo has an information point.
- Visitors use a menu to select information about animals.
- The menu includes 500 different animals.
- The information is provided only using high definition video with an audio track.

a State **one** input device that could be used for the information point. [1]

b The output is shown on a monitor.
State **one** other output device that could be used for the information point. [1]

c The video files are stored at the information point.
State **one** secondary storage device that could be used. [1]

d The zoo decides to introduce Quick Response codes in different places in the zoo. These provide further information about the animals.
Describe how customers obtain the information from the Quick Response codes. [4]

Cambridge IGCSE Computer Science 0478, Paper 11 Q4, Oct/Nov 2019

8 Anna has a farm that grows fruit.
She has a system that monitors the conditions for growing the fruit. Sensors are used in this system.

a Explain what is meant by the term **sensor**. [2]

b State **two** sensors that could be used in this system and describe how they could be used. [6]

Cambridge O Level Computer Science 2210, Paper 12 Q9, Oct/Nov 2017

9 The diagram shows **five** output devices and **five** descriptions.
Draw a line between each output device and its description. [4]

Output device	Description
Inkjet printer	Flat panel display that uses the light modulating properties of liquid crystals.
LCD screen	Flat panel display that uses an array of light emitting diodes as pixels.
2D cutter	Droplets of ink are propelled onto paper.
LED screen	Electrically charged powdered ink is transferred onto paper.
Laser printer	2D High-powered laser that uses the X-Y plane. cutter

Cambridge O Level Computer Science 0478, Paper 12 Q3, May/June 2017